Realising Phase Imbalance Assessments and Phase Swapping for Data-Scarce Low Voltage Networks
(Alternative Format Thesis)

Fang, Lurui

Award date:
2021

Awarding institution:
University of Bath

UNIVERSITY OF
BATH

Realising Phase Imbalance Assessments and Phase Swapping for Data-Scarce Low Voltage Networks

By

Lurui Fang

BEng, Msc

The thesis submitted for the degree of

Doctor of Philosophy

in

The Department of
Electronic and Electrical Engineering
University of Bath

August 2021

Signature:…… Date:……

Contents

Abstract

Phase imbalance is a widespread and long-outstanding problem for low voltage (LV, 415V) networks. According to the data from Western Power Distribution (WPD, one of the UK's distribution network operators), more than 50% of their LV networks suffered from severe phase imbalance – the current on the "heaviest" phase exceeded that on the "lightest" phase by 50% most of the time, increasing energy losses by up to 30%. Moreover, phase imbalance leads to inefficient use of network assets. If leaving phase imbalance unsolved, distribution network operators have to reinforce LV networks when the "heaviest" phase goes overloaded, despite having unused capacities on the "lightest" phase. Suppose an LV network suffers from significant phase imbalance where its "heaviest" phase exceeds the "lightest" phase by 50%, and its "heaviest" phase uses up the capacity of that phase. If leaving phase imbalance as it is, the LV network requires reinforcement immediately. However, supposing the annual load growth rate is 2%, fully rebalancing this LV network can defer network reinforcements for at least 15 years.

A number of references had studied imbalance-induced consequences and phase balancing solutions, supposing they had substation-side time-series phase current data and customer-side smart meter data. However, in reality, those data are not collected due to the lack of advanced monitoring devices for the majority of LV networks in the UK. It, therefore, raises a solid challenge: distribution network operators cannot implement existing phase imbalance assessments and phase swapping to over 900,000 LV networks in the UK because of data scarcity. This thesis, for the first time, addresses this challenge by developing statistical methodologies to estimate imbalance-induced energy losses and making phase swapping (one classic phase balancing method by moving customers from one phase to the other) guidance for data-scarce LV networks. Compared to existing solutions, this thesis' developed

methodologies only require existing data from data-scarce LV networks, thus accommodating the reality and being practical for industrial implementation.

This thesis originally develops three methodologies to get around the data limitations in implementing phase imbalance diagnosis and phase swapping. First, this thesis estimates imbalance-induced energy losses on the phase residual path and the three phases, separately, using only yearly average and maximum phase current data. The estimation accuracies are over 80% and 82%, respectively. Second, this thesis develops phase swapping guidance for LV networks without the requirement of year-round substation-side time-series phase current data and customer-side smart meter data. Case studies reveal that my approach achieves effective reductions of the phase imbalance degrees for data-scarce networks. And the reduction of phase imbalance degree is only 14.3% lower than that for data-rich networks.

Given the developed approaches, this thesis brings the following solid contributions for the industry: 1) it provides one significant component of the decision making of phase balancing investments, imbalance-induced energy losses, for data-scarce LV networks; and 2) it significantly improves the practically of phase swapping for the industrial implementation. In short, this thesis turns massive industrial application of the decision making for phase balancing investments and phase swapping planning into reality.

Publications

1. L. Fang, K. Ma and X. Zhang, “A Statistical Approach to Guide Phase Swapping for Data-Scarce Low Voltage Networks,” IEEE Transactions on Power Systems vol. 35, no. 1, pp. 751-761, Jan. 2020, doi: 10.1109/TPWRS.2019.2931981.

2. L. Fang, K. Ma, R. Li, Z. Wang and H. Shi, “A Statistical Approach to Estimate Imbalance-Induced Energy Losses for Data-Scarce Low Voltage Networks,” IEEE Transactions on Power Systems vol. 34, no. 4, pp. 2825-2835, July 2019, doi: 10.1109/TPWRS.2019.2891963.

3. L. Fang and K. Ma, “Assessment of additional phase energy losses caused by phase imbalance for data-scarce LV networks,” IET Generation, Transmission & Distribution, vol. 14, no. 4, pp. 675-681, 28 2 2020, doi: 10.1049/iet-gtd.2019.1036.

4. L. Fang, K. Ma, and Z. Zhong, ” A Statistical Methodology to identify Imbalance-Induced Capacity Wastes for LV Networks”, Electric Power System Research, under first-round revision.

5. K. Ma, L. Fang and W. Kong, "Review of distribution network phase unbalance: Scale, causes, consequences, solutions, and future research directions," CSEE Jmynal of Power and Energy Systems, vol. 6, no. 3, pp. 479-488, Sept. 2020, doi: 10.17775/CSEEJPES.2019.03280.

6. W. Kong, K. Ma, L. Fang, R. Wei and F. Li, "Cost-Benefit Analysis of Phase Balancing Solution for Data-Scarce LV Networks by Cluster-Wise Gaussian Process Regression," in IEEE Transactions on Power Systems, vol. 35, no. 4, pp. 3170-3180, July 2020, doi: 10.1109/TPWRS.2020.2966601

Acknowledgement

Firstly, I would like to express my deepest gratitude to my supervision team, Dr Kang Ma, Prof. Furong Li, and Dr Ran Li, for their invaluable guidance, support and inspiration throughout my research. I appreciate becoming a member of the CSPD family.

I would like to express my thanks to University Bath, who funded my PhD research. I would thank everyone who supports my work from the Department of Electrical and Electronic Engineering.

I would like to express my gratefulness to Dr Chenghong Gu, Dr Xinsong Zhang, Dr Yunjie Gu, Dr Xiaoze Pei, Dr Ignacio Hernando Gil, and Dr Zhong Zhang for constructive suggestions and advice. I would also like to express my thanks to my fellow friends, Dr Heng Shi, Dr Zhong Zhang, Dr Xinsong Zhang, Dr Han Wu, Dr Wenjuan Song, Dr Wangwei Kong, Mr Haiwen Qin, and Mr Renjie Wei, who give me strong supports on my research and settling in Bath.

I would like to express my heartfelt gratitude to Prof. George Chen and Mrs Ying Wang, who consistently inspire and guide me down the right path.

And I would like to take this opportunity to express my ultimate thanks to my wife, Xiaoshan, who nearly dedicate everything to support my dream and pursuit. I would also thank my parents in law for their invaluable support in pursuing my dream. Last but not least, I would like to express my gratitude to my parents, Xueming and Mingling, for their love and encouragement, without whom I would never be such close to the world-leading research.

List of Figures

List of Tables

List of Abbreviations

LV	Low Voltage
WPD	Western Power Distribution
SPEN	Scottish Power Energy Networks
DNO	Distribution Network Operator
LCT	Low Carbon Technology
HP	Heat Pump
EV	Electric Vehicle
ARC	Additional Reinforcement Cost
DG	Distributed Generator
DER	Distributed Energy Resource
APEL	Additional Phase Energy Loss
RMS	Root-Mean-Square
DIB	Degree of Phase Imbalance
RTU	Remote Telemetry Unit
CCRE	Clustering, Classification, and Range Estimation
CDF	Cumulative Density Function
SVR	Support Vector Machine Regression
MSVM	Multiclass Support Vector Machine
IIBL	Imbalance-Induced Energy Loss
RMSE	Root-Mean-Squared Error
ED	Euclidean Distance
JSD	Jensen-Shannon Distance
AMM	Automated meter management
NMF	Non-Negative Matrix Factorisation
RP	Rebalancing Potential

Chapter 1.

Introduction

Chapter contents:

This chapter overviews the background, motivation, objectives, challenges and contributions. It also presents the structure of this thesis.

1.1. Background and motivation

1.1.1. Drives: phase imbalance increases the operation costs of low voltage networks

In the UK's low voltage (LV, 400V) networks, the majority of terminal customers are single-phase connected to the grid. This inevitably causes two circumstances that both incur phase imbalance for LV networks. First, active customers on the network's three phases are normally not in the same number because of house ownership changes, network maintenance and modification. Second, customers' load profiles vary from customer to customer throughout a day. Unbalanced customer quantities and their different load profiles lead to different phase load profiles for the LV network. The difference among the phase load profiles calls phase imbalance. According to the data from Western Power Distribution (WPD), one of the UK's distribution network operators (DNOs), more than 50% of its LV networks suffered from severe phase imbalance. Their "heaviest" phase current exceeded the "lightest" phase current by 50% most of the time [1]. TNEI, a UK consultancy, found that 165 of 233 (more than 70%) of Scottish Power Energy Networks' (SPEN) LV lines had noticeable phase imbalance, where the greatest single-phase current exceeded the average phase current by 30% most of the time [2], [3]. In the future, phase imbalance will worsen when low carbon technologies (LCTs) are randomly and unorderly connected to the three phases of LV networks [2], such as electric vehicles (EVs), heat pumps (HPs), household PV systems. By 2030, UK power networks estimate that there could be over 700,000 electric heat pumps and 4.5 million EVs across London, the East and South East of the UK [4].

Phase imbalance is also noticeable in other countries' LV networks. For example, reference [5] indicated phase imbalance is a problem for Denmark's LV networks,

where customers were three-phase connected to the LV networks. Reference [6] presented 2% of MV networks in the US had undesirable phase imbalance.

Phase imbalance leads to a number of consequences:

- Insufficient infrastructure utilisation – imbalance-induced capacity wastes. If leaving phase imbalance unsolved, DNOs have to reinforce LV networks when the “heaviest” phase goes overloaded, despite having unused capacity on the other two phases.

- Additional energy losses [7], [8] on feeders, transformers, and the residual path. Totally, phase imbalance raises energy losses by up to 35% [1].

- Substantial zero-sequences currents. Unexpected zero-sequence currents could cause malfunction of relay protection devices (e.g. the zero-sequence overcurrent relays for transformers), increasing network tripping risks [9]. According to the data from WPD, the zero-sequence currents could achieve up to 2 times the ‘lightest’ phase current.

- Motor overheating and damage from severe voltage imbalance [10], [11].

Overall, phase imbalance increases LV networks’ operation costs [7], [8], [12], especially when the bulk of LCTs are randomly and unorderly connected to the grid in the foreseeable future [2]. To minimise this cost, having a credible phase imbalance diagnosis is vital for DNOs to make phase balancing investment decisions.

1.1.2. Challenges: data scarcity, scalability and inconvenience

A number of references had studied phase imbalance consequences and phase balancing solutions. However, massively implementing them on real networks faces limitations.

- **The data-scarcity problem for LV networks**

Theoretically, assessing phase imbalance consequences and making phase balancing decisions is a straightforward engineering task. Adopting time-series phase current data at the substation side of LV networks can directly drive imbalance-induced energy losses [8], [13], [14] and imbalance-induced capacity wastes (expressed by additional reinforcement costs [12], [15]), alongside determining phase balancing solutions. However, only a small portion of the UK's networks collect these year-round time-series phase current data in reality. The majority of LV networks in the UK do not record those data due to a lack of advanced monitoring devices. The data-scarcity problem makes existing phase imbalance assessment methodologies not applicable to the majority of LV networks and limits the implementation of off-line phase balancing solutions, such as phase swapping and network reconfiguration.

- **The scalability and inconvenience problems for existing phase imbalance assessments and phase balancing solutions**

To address the data scarcity problem, one solution is deploying advanced monitoring devices. However, implementing this solution for over 900,000 LV networks incurs prohibitive costs, thus incurring practicality and scalability problems.

Moreover, apart from the data-scarce problem, existing phase balancing solutions also perform other scalability and practicality limitations. The state of the art phase balancing solutions have three categories: 1) off-line phase balancing solutions, such as phase swapping and network reconfiguration [16], [9], [17]; 2) tailor-designed phase balancing systems, such as customer-side phase switches [18], [19], [20] and power-electronics-based phase balancers [21], [22], [23], [24]; and 3) using existing LCT devices to provide phase balancing [7], [22], [25], [26].

First, off-line balancing solutions require scheduled power cuts and intensive fieldworks, as well as network topologies that are not properly documented by DNOs in the UK [27], [28]. These requirements limit the scalability for off-line balancing solutions. Furthermore, off-line balancing solutions cannot guarantee long-term phase balancing effectiveness. Unbalanced load changes, particularly the random connection of single-phase LCTs, would change the imbalance direction from time to time, invalidating previous off-line balancing strategies. It also implies DNOs require deploying off-line phase balancing multiple times to guarantee long-term phase balancing effect.

Second, deploying phase switches incurs prohibitive costs. For example, it requires deploying at least 27.8 million switches in the same number of households in the UK [29]. Furthermore, deploying phase switches requires excavating roads and modifying link boxes on cables. Millions of phase switches bring vastly inconvenience for DNOs when maintenance is required. For example, if the annual failure rate for phase switches is 0.01%, it requires maintenance works for at least 2,780 phase switches each year in the UK. It should be noted that the failure rate increases with the phase switch ageing.

Third, using existing LCT devices, such as energy storage systems, EV chargers, and household PV systems, can theoretically address phase imbalance. However, a gap remains between the technical solution and real business implementation: there is currently no business approach to motivate these customers to provide their LCT devices for phase balancing, especially guiding customers to prioritise predominant imbalance-induced consequences.

To sum up, there raises the following challenges: 1) assessing phase imbalance consequences and making phase balancing decisions for data-scarce LV networks; and 2) subject to 1), those solutions should be tangible, scalable and practical.

1.2. Research contributions

This thesis makes original contributions by two means: 1) for the first time developed phase imbalance assessment approaches for data-scarce LV networks (LV networks that do not record substation-side year-round time-series phase current data); 2) for the first time developed phase swapping (one of the off-line phase balancing solutions by swapping customers from one phase to the other) guidance for data-scarce LV networks.

- **Assessing imbalance-induced energy losses for data-scarce LV networks**

Phase imbalance leads to imbalance-induced energy losses, including losses on the three phases and the phase residual path. According to these two consequences, this thesis develops two corresponding statistical methodologies to get around the daunting data-scarce problem for LV networks. These approaches do not require any deployment of additional monitoring devices, thus providing a cost-effective way to implement phase imbalance assessment massively in the industry.

First, this thesis develops a regression-based approach to assess the imbalance-induced phase energy losses for data-scarce LV networks. Unlike existing phase imbalance assessment studies, this approach only requires the yearly average and maximum phase currents, thus applicable to the majority of LV networks. Through validation in Chapter 3, this approach delivers at least 80.6% estimation accuracy for 90% of data-scarce LV networks within WPD's business area.

Second, this thesis develops a clustering, classification & range estimation (CCRE) approach to assess the imbalance-induced residual energy losses for data-scarce LV networks. Similar to the regression-based approach, this approach only requires the yearly average phase currents, thus applicable to the majority of LV networks.

Compared to the regression-based approach, which delivers a point data estimation, the CCRE approach derives range estimations. Given that phase imbalance changes are uncertain in the long term, range estimations are more credible than point estimation for DNOs in making phase balancing investment decisions.

- **Developing phase swapping guidance for data-scarce LV networks**

Phase swapping is a cost-effective phase balancing solution by moving single-phase customers from one phase to the other [2]. Previous references developed phase swapping strategies based on detailed network topologies, time-series phase loading data, and customer's smart meter data. However, implementing these methods on real networks faces a challenge: the majority of the UK's LV networks do not have time-series phase load data and customer's smart meter data. In deploying phase swapping for a mass scale of LV networks, the data-scarcity problem offsets the cost-effective advantage of phase swapping. To address this challenge and promote the scalability of phase swapping, this thesis originally develops a statistical approach to guide phase swapping for data-scarce LV networks without the requirement of customer-side smart meter data and year-round substation-side phase current data. Case studies in Chapter 5 reveal that the statistical approach produces effective phase swapping guidance, which reduces the phase imbalance degrees for 99% of the LV networks. The maximum reduction of phase imbalance degrees is 0.35. Moreover, the reduction of phase imbalance degree for data-scarce LV networks is only 14.3% lower than that for data-rich networks.

1.3. Thesis layout

The rest of the thesis is organised as follows:

Chapter 2 reviews the existing literature on phase imbalance consequence

assessment and phase balancing solutions, including the state of art studies and industry adopted approaches.

Chapter 3 estimates imbalance-induced phase energy losses for data-scarce LV networks. It develops one novel statistical approach. This approach only requires existing data from data-scarce LV networks, and delivers over 80% estimation accuracy in estimating imbalance-induced phase energy losses.

Chapter 4 estimates imbalance induced residual energy losses for data-scarce LV networks. It develops a clustering classification and range estimation approach. This approach not only delivers over 80% accuracy but also perform a range estimation. This range estimation is more valuable for DNOs to make long-term phase balancing investment decisions than point estimation.

Chapter 5 derives phase swapping guidance for data-scarce LV networks by three steps. To achieve that, this thesis develops a statistical approach on top of data analytics. This approach outputs the number and the type of loads that are required to swap from a given “heavy” phase to one “light” phase. For example, the derived guidance is “swapping 3 commercial loads from heavy-loaded phase A to light-loaded phase B” to mitigate phase imbalance. This chapter also discusses the implementation of the derived phase swapping guidance for data-scarce LV networks.

Chapter 6 performs conclusions and discusses potential future works.

Chapter 2.

Review of phase imbalance and phase balancing solutions

Chapter contents:

This chapter reviews the literature of phase imbalance assessments and phase balancing solutions. It includes the state of art studies and industry adopted approaches.

2.1. Introduction

Phase imbalance is a widespread and long-outstanding problem for LV networks around the world [30], [31], [32]. This section makes a comprehensive review of phase imbalance, impacts & assessments, and phase balancing solutions. The review includes state-of-the-art studies and industry adopted approaches.

2.2. Phase imbalance for LV networks

Typically, phase imbalance includes two scenarios: load imbalance among the three phases and voltage imbalance among the three phases.

2.2.1.Load imbalance

For LV networks within UK, China and substantial countries, customers are single-phase connected to the grid. The uneven number of customers and their vastly different behaviour inevitably lead to different loading levels on the three phases on the time horizon. This difference is load imbalance. The load imbalance severities from different countries are shown in Fig.2-1. For the two of UK's DNOs, WPD and SPEN, 50% and 70% of LV networks have siginifcant phase imbalance problems, where the current on the "heaviest" phase exceeds that on the "lightest" phase by 50% for the majority of time.

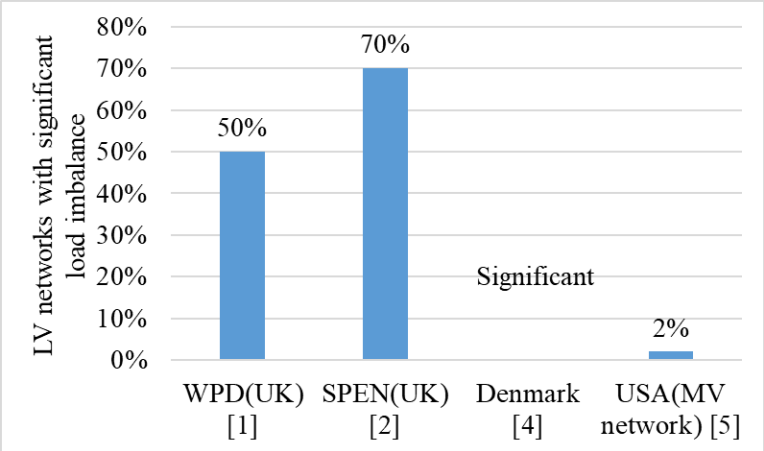


Fig. 2-1 Load imbalance severity in different countries

Since this paper focuses on phase imbalance (load imbalance) assessment and phase swapping decision making, the details of load imbalance consequences and load balancing solutions are presented in Chapters 2.3 and 2.4.

2.2.2. Voltage unbalance

Different from the principle of load imbalance, voltage imbalance includes unequal voltage values, phase angle deviation and different harmonic distortion levels on the three phases. The national electrical manufacturers association (NEMA) defined voltage unbalance as the ratio of the maximum voltage variation to the average three-phase voltage [33], as given by:

$$VUF = \frac{\max\{V_a, V_b, V_c\} - ave\{V_a, V_b, V_c\}}{ave\{V_a, V_b, V_c\}} \quad 2\text{-}(1)$$

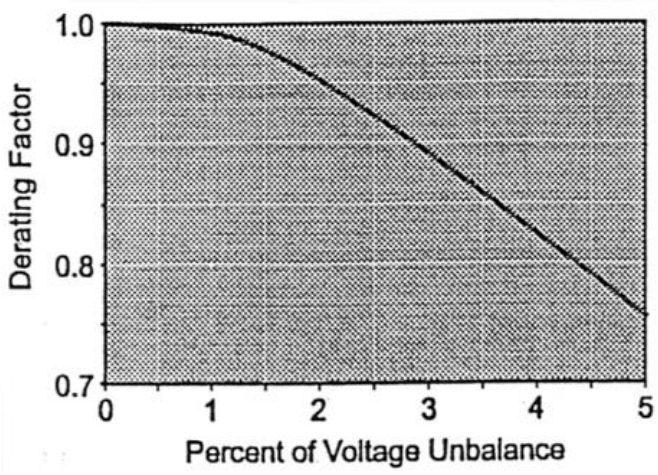
where $ave\{\quad\}$ is the function for calculating the average value of all variables within the brace; V_a, V_b, V_c denote the voltage on phase a, b, and c, respectively. According to the data from U.S. distribution systems, 66% of distribution networks had a voltage unbalance degree of up to 1%. For 32% of distribution networks, the voltage unbalance degree was between 1% and 3% [34].

Further, voltage unbalance has a second definition, which gives a ratio of the negative sequence component of the three-phase voltage to the positive sequence component of the three phases.

$$VUF = V^{-}/V^{+} \tag{2-(2)}$$

where V^{-} and V^{+} are the negative sequence component of voltage to the positive sequence component of voltage, respectively.

Voltage unbalance leads to two main consequences: derating and additional energy losses for induction motors, including customer-side three-phase motors and network-side transformers [35]. The derating and energy loss increase problems are presented in Fig. 2-2 and Fig.2-3. 5% of volatge unbalance leads to over 20% derating and 15% of energy loss increase for induction motors.



Voltage Unbalance	Approximate Derating
1%	None
2%	95%
3%	88%
4%	82%
5%	75%

For motors rated greater than five-hundred horsepower, consult the vendor about derating.

Fig. 2-2 Voltage unbalance-induced derating for induction motors [36]

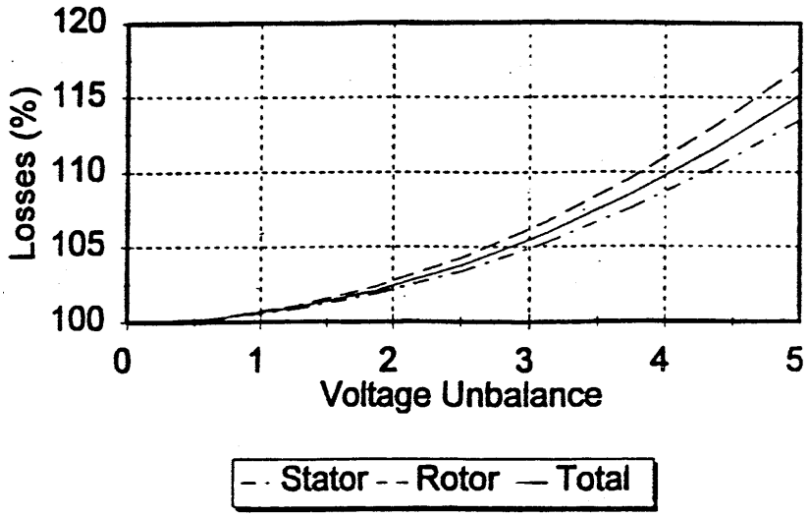


Fig. 2-3 The voltage-unbalance-induced energy loss increase for 240V 25 hp induction motors [37]

Typically, load imbalance and voltage unbalance interact with each other. The main cause for voltage unbalance is uneven loading levels on the three phases [36], while voltage unbalance also siginificantly exacerbates load imbalance [35]. This thesis studies the load imbalance problem for LV networks, including its real problem and solutions.

2.3. Phase imbalance consequences and assessments

Phase imbalance (load imbalance) leads to two major consequences: additional energy losses and capacity wastes. This section summarises existing approaches for assessing these two consequences.

2.3.1. Imbalance-induced energy loss assessments

Phase imbalance leads to two types of additional energy losses. First, phase imbalance increases energy losses on phase wires [8]. Second, it generates zero sequence current, which causes additional energy losses on the residual path [8]. For LV networks with TN-C earthing systems, shown in Fig. 2-5, the residual path is the

ground between customers and the transformer [38]. For LV networks with TN-S earthing systems, shown in Fig. 2-4, the residual path is the neutral wire between customers and the transformer [38].

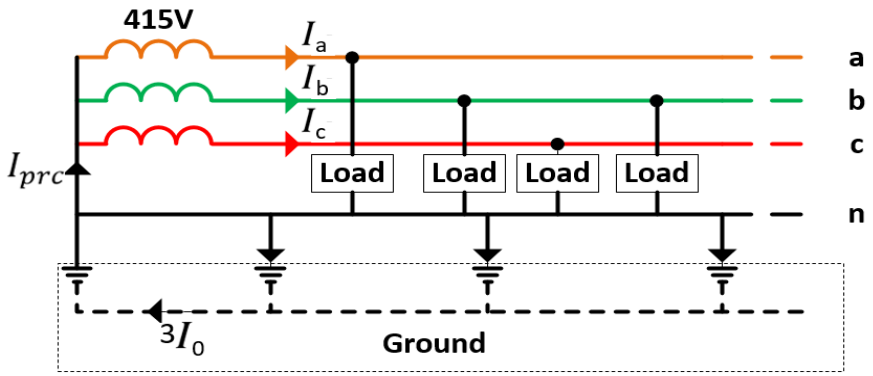


Fig. 2-4 Examples for TN-C earthing systems

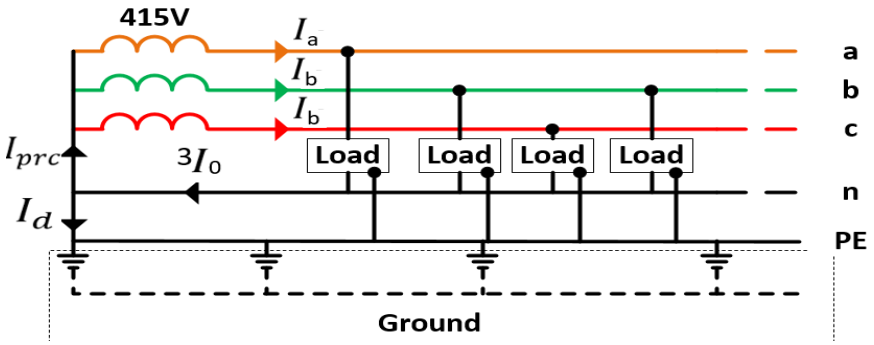


Fig. 2-5 Examples for TN-S earthing systems

Theoretically, the imbalance-induced energy loss is given by:

$$E_{IB} = \left(\sum_t I_{a,t}^2 R_p + I_{b,t}^2 R_p + I_{c,t}^2 R_p - \sum_t \left(\frac{I_{a,t} + I_{b,t} + I_{c,t}}{3}\right)^2 R_p\right) + \sum_t {I_{pr}}^2 R_{rp} \qquad 2\text{-}(3)$$

where $I_{pr} = \sqrt{I_a^2 + I_b^2 + I_c^2 - I_a I_b - I_b I_c - I_a I_c}$;

$I_{a,t}$, $I_{b,t}$, and $I_{c,t}$ are currents for phase a, b, and c, respectively, at time t; R_p is the equivalent resistance for phases; I_{pr} is the phase residual current; R_{rp} is the equivalent resistance for the residual path. Equation 2-(3) is adopted only when time-series phase current data are collected in a given period.

According to data from WPD [1], Fig. 2-6 shows the percentage of imbalance-induced energy losses, where the utility price is 0.18 £/kWh.

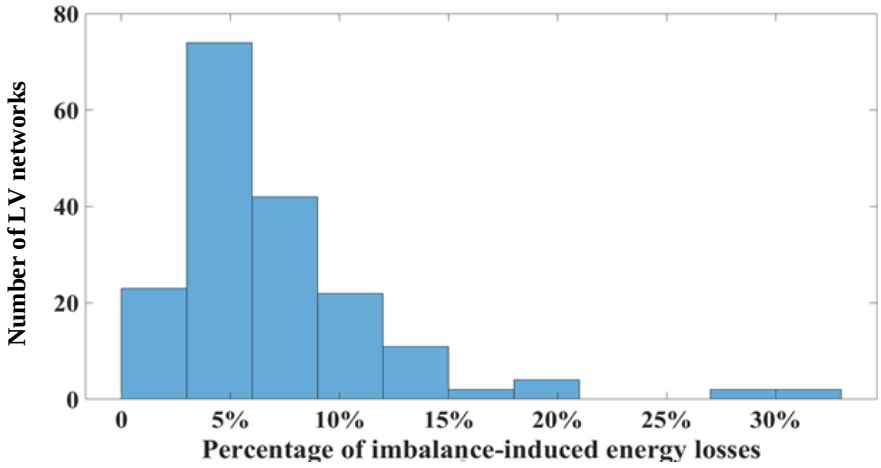


Fig. 2-6 The percentage of imbalance-induced energy losses for LV networks within WPD's business area

The rest of the section summarises research studies for calculating and estimating imbalance-induced energy losses, as well as the industry-adopted approaches for estimating energy losses.

- Research studies for calculating imbalance-induced energy losses

Reference [8] calculated imbalance-induced energy losses for 29-bus and 34-bus example networks using detailed power flow analysis. Reference [13] modelled the line segments by Carson's equation to calculate the neutral energy losses (imbalance-induced residual losses) of overhead lines. Reference [14] defined a ratio between the equivalent neutral line resistance and the phase wire resistance to estimate neutral energy losses. Reference [39] calculated the neutral energy losses induced by non-linear three-phase loads. Reference [40] calculates imbalance-induced additional copper losses for LV transformers. Reference [41] formulates a complex unbalance factor for calculating imbalance-induced energy losses for LV networks.

- Research studies for estimating energy losses

Because of inadequate metering in LV networks, the energy losses are normally estimated instead of measured. Reference [42] developed two energy estimation schemes: 1) using phase current data to estimate energy losses for MV networks, and 2) using integrating type metering data to estimate energy losses for LV networks. Reference [43] developed a simplified unbalanced power flow model to estimate energy losses.

- The industry adopted approaches to estimate energy losses

The industry commonly adopted the load loss factor to estimate energy losses [44]. This approach derives the relationship between the load factor and loss factor to estimate the average energy loss in a given period. The following steps describe the load loss factor model:

1) First, the load factor is calculated using the load profiling data throughout a given period.

$$Load\ factor = Load_{avg}/Load_{peak} \quad 2\text{-}(4)$$

where $Load_{avg}$ is the average loading level throughout a given period, such as one month; $Load_{peak}$ denotes the peak loading level throughout a given period.

2) Second, the loss factor is calculated based on the calculated load factor.

$$Loss\ factor = Loss_{avg}/Loss_{peak} \quad 2\text{-}(5)$$

where $Loss_{avg}$ is the average energy losses throughout a given period, such as one month; $Loss_{peak}$ denotes the peak energy losses throughout a given period.

3) Third, according to calculated load factor and loss factor, a weight factor a is estimated, as given by:

$$a = \frac{Loss\ factor - (Load\ factor)^2}{Load\ factor - (Load\ factor)^2} \quad 2\text{-}(6)$$

4) Given the derived a and the $Load\ factor$, equation 2-(6), in turn, could estimate the loss factor for the future. By for estimating future energy losses, the average energy loss is equal to the peak energy loss multiplying the estimated loss factor.

It should be stressed that although the load loss factor (LLF) model is the typical industry-adopted method for estimating energy losses. However, it also has limitations: DNOs require updating the LLF model each month to obtain an accurate estimation [44].

2.3.2. Imbalance-induced capacity waste assessments

Phase imbalance leads to insufficient use of the LV network's infrastructure. DNOs require reinforcing LV networks when the "heaviest" phase goes overloaded, despiting having unused capacity on "light" phases. This premature network reinforcement brings additional costs. Fig. 2-7 presents an example of imbalance-induced capacity wastes where the imbalance-induced capacity waste is 33A.

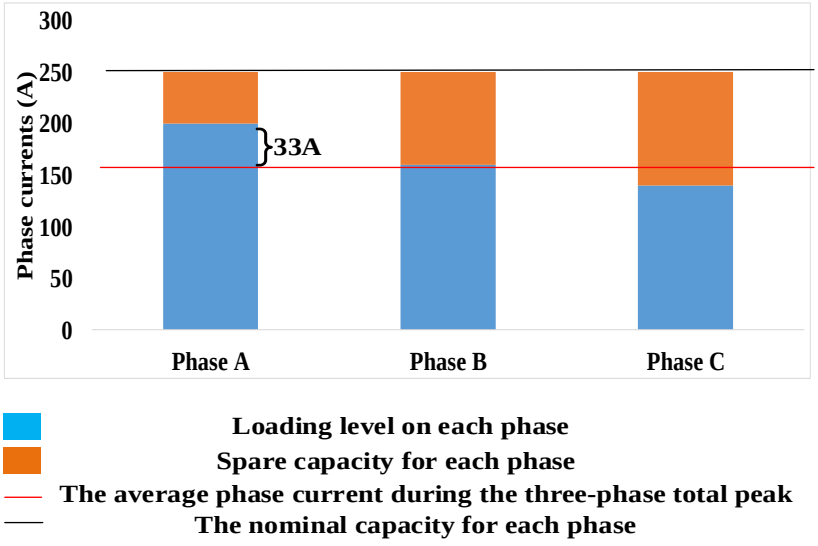


Fig. 2-7 An example of the imbalance-induced capacity wastes

Previous literature studies imbalance-induced capacity wastes in two ways: 1)

assessing imbalance-induced capacity wastes in kW [2], and 2) converting the kW capacity wastes into additional reinforcement costs (ARCs) [45] [12].

- Assessing imbalance-induced capacity wastes in kW

Reference [2] studied imbalance-induced capacity wastes for LV networks within Scottish Power Energy Networks' operation area. The potential capacity release via phase balancing is given by:

$$r_{cl} = \frac{I_{fh} - I_{sh}}{I_{fh}} \quad 2\text{-}(7)$$

Where I_{fh} is the peak current on the heaviest phase, I_{sh} is the peak current of the second-highest phase. It should be noted that equation 2-(7) does not calculate the maximum potential release. Reference [2] noted that r_{cl} is the potential release with minimal phase balancing costs.

Overall, data from SPEN showed that 165 out of 233 feeders had an average phase imbalance ratio of no less than 1.3. This implies that the "heaviest" phase current exceeds the average phase current by 30% most of the time. Approximately 10% of these feeders had the potential to release more than 20% capacity headroom, which equals no less than 100A capacity.

- Converting Amps capacity wastes to additional network reinforcement costs

Given that DNOs will reinforce LV networks when the "heaviest" phase goes overloaded, phase imbalance leads to premature network reinforcements. Deploying phase balancing solutions can defer network reinforcements for years to save investments. By contrast, leaving phase imbalance unsolved incurs additional reinforcement costs.

For example, suppose 1) the DNO requires investing £10,000 to reinforce an LV

network because of single-phase overload; 2) the DNO can defer network reinforcements for ten years if this LV network is fully rebalanced. If the three phases are balancing, the DNO's stakeholders can invest these £10,000 to other areas at present and make profits over the next ten years. Suppose the rate of return is 6.9% for other investments and the yearly inflation rate is 2% over the ten years. This implies DNO's stakeholders can make profits of £9,500, where the net present profits are £8,400, via investing £10,000 in other areas. However, because of phase imbalance, DNOs must invest that money in network reinforcements to address network congestions today, thus losing the calculated profits over the next ten years. The lost profits are the ARC.

Reference [12] develops a formula to quantify phase imbalance to ARCs. The ARC is given by:

$$ARC = C(1+d)^{\frac{\log U_{tp}}{\log(1+r)}}\left[(1+d)^{\frac{\log(3D_{tp}+1)}{\log(1+r)}} - 1\right] \quad 2\text{-}(8)$$

where C is the reinforcement costs for the network in question; U_{tp} is the utilisation rate during the three-phase total peak; D_{tp} is the phase imbalance degree during the three-phase total peak; $P_{\varphi}(t)$ is the current on phase φ at time t, $\varphi\epsilon\{a,b,c\}$; r is the load growth rate; d is the discount rate.

According to data from WPD [1], Fig. 2-8 shows the imbalance-induced additional reinforcement costs, where the network reinforcement costs are the same as those in TABLE 5-5 and TABLE 5-6.

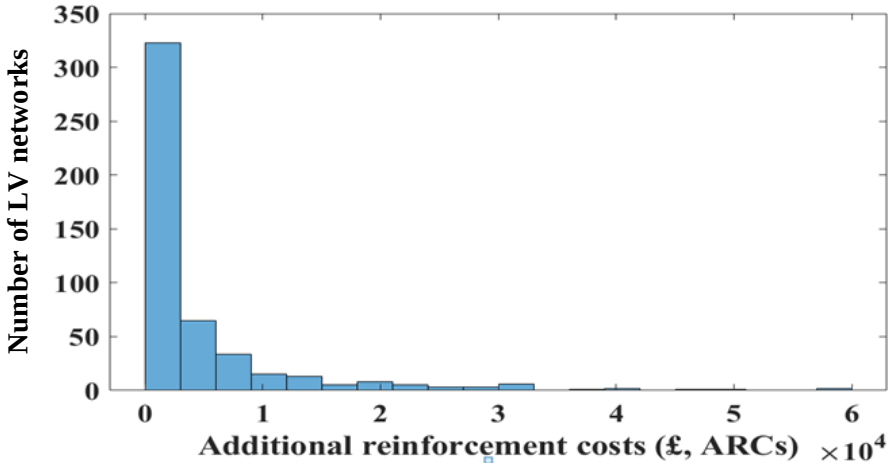


Fig. 2-8 Imbalance-induced capacity wastes for LV networks within WPD's business area

2.4. Phase balancing solutions

Existing phase balancing solutions have three categories: 1) off-line phase balancing solutions; 2) tailor-designed phase balancing systems, such as power-electronics-based phase balancers and deploying customer-side phase switches; and 3) using LCT devices to provide phase balancing.

- Off-line phase balancing solutions

Phase swapping and network reconfiguration are two classic off-line phase balancing solutions. DNOs implement them during scheduled power cuts by moving customers from one phase to the other. It requires detailed network topology, historical substation-side time-series phase currents data, and customer-side smart meter data to plan off-line phase balancing. Although this solution is classic, its implementation difficulty varies from case to case. For overhead lines, field engineers can quickly locate the joints between customers and the main feeder for implementing phase swapping. For underground cables, field engineers require excavating roads and make new joints on cables. Moreover, off-line phase balancing cannot ensure a long-term phase balancing

effectiveness – any phase imbalance direction change in the future invalidates previous off-line phase balancing solutions.

A number of references had studied phase swapping or network reconfiguration guidance. This guidance presents how to swap customers to achieve maximal phase balancing effectiveness. However, combining the principle of off-line phase balancing solutions and power flow analysis creates a complex optimisation problem. This optimisation problem is given as follows:

To address this problem, 1) reference [16] used simulated annealing to formulate an optimal power flow model; 2) reference [9] formulated a mixed-integer optimisation model, where both customer and lateral swapping were considered; 3) references [46] and [47] developed a combined heuristic and neural network model and a combined fuzzy logic and Newton-Raphson model, respectively; reference [48] developed a novel dynamic programming algorithm; reference [49] formulated a genetic algorithm considering a detailed power flow model; reference [50] developed a backward sweep model for DG-Integrated distribution networks. Furthermore, reference [51] developed a hierarchical decentralised approach to make network reconfiguration strategy for unbalanced distribution networks with distributed generations. Reference [52] developed a fuzzy evolutionary particle swarm optimisation to solve phase swapping optimisation problems.

In the implemenation, off-line phase balancing solutions performs obvious limitations. they require scheduled power cuts and intensive fieldworks, as well as network topologies that are not properly documented by DNOs in the UK [27], [28]. Furthermore, off-line balancing solutions cannot guarantee long-term phase balancing effectiveness. Unbalanced load changes, particularly the random connection of single-phase LCTs, would change the imbalance direction from time to time, invalidating previous off-line balancing strategies.

- Tailor-designed phase balancing systems

Tailor-designed phase balancing systems have two categories: 1) power-electronics-based phase balancer and 2) deploying phase switches for each customer.

The principle of power-electronics-based phase balancers is a controllable bridge among the three phases that transfers loads from “heavy” phases to “light” phases. This principle determines power-electronics-based phase balancers only address the phase imbalance upstream, while the downstream phase imbalance remains unsolved. Its deployment location, therefore, is vital to planning. Reference [53] developed a three-phase electric spring circuit for phase balancing. Reference [54] used D-STATCOM for phase balancing. Reference [21] reallocated the three-phase currents of the AC/DC converter, located between AC and DC networks, for phase balancing. References [23] and [55] developed-static-var compensator-based phase balancer to eliminate negative and zero sequence currents. Reference [24] developed an active phase filter to inject negative currents into the grid for phase balancing.

The principle of using phase switches for phase balancing is the same as the off-line phase balancing solutions. Using phase switches for phase balancing addresses the limitation of not accommodating future phase imbalance changes for offline phase balancing solutions. However, its implementation costs and maintenance inconvenience problems should be further investigated. References [18], [19], [20] studied how to switch customers for phase balancing. Reference [20] combined phase switch control with phase identification to improve the practicality of deploying phase switches for phase balancing. Reference [56] formulated phase switching guidance by considering customer’s load profiles. Reference [57] formulated a dynamic and heuristic model to derive phase switching guidance

However, deploying phase switches incurs prohibitive costs. For example, it requires

excavating roads modifying link boxes and deploying at least 27.8 million switches in the same number of households in the UK [29]. Further, millions of phase switches bring vastly inconvenience for DNOs when maintenance is required. For example, if the annual failure rate for phase switches is 0.01%, it requires maintenance works for at least 2,780 phase switches each year in the UK. It should be noted that the failure rate increases with the phase switch ageing.

- Using LCT devices to provide phase balancing

Using LCT devices, such as energy storage systems, EV chargers, and household PV systems, for phase balancing breaks down into two categories: 1) reallocating phase currents of three-phase LCT devices for phase balancing; 2) shifting single-phase LCT loads on time horizons for phase balancing.

First, references [58] and [59] justified that three-phase AC/DC converters can operate under the unbalanced model to provide phase balancing for LV networks. For example, suppose the three-phase currents at time *t* are [10A, 20A, 30A] for an LV network, a three-phase LCT device's load is 30A loads at time *t*, and the single-phase capacity for this LCT device is 20A. If the LCT does not provide phase balancing, its three-phase load current is [10A, 10A, 10A]. If it provides phase balancing, the control will reallocate its three-phase load current to [20A, 10A, 0A]. At this time, the LCT's AC/DC converter works under an unbalanced model, while the three-phase currents for the LV network are rebalanced to [20A, 20A, 20A]. Based on the principle of [58] and [59], reference [21] used the AC/DC converters between the AC and DC networks for phase balancing. Reference [22] used the spare capacities of three-phase PV systems and EV chargers' converters for phase balancing. Reference [60] adopted a transactive approach to coordinate the operations of DERs and data centres for phase balancing. This approach was applied in the distribution electricity market to justify its profits. Reference [61] controlled the inverters of single-phase distributed generations to offset

negative and zero sequences current for LV networks, thus achieving phase balancing.

Second, when there are no distinguished “heavy” phases and “light” phases on time horizons, shifting single-phase LCT loads on time horizons for phase balancing can address phase imbalance. The “no distinguished “heavy” phases and “light” phases” indicates each phase will alternately become the “heavy” phase or “light” phase on the time horizon. Fig. 2-9 presents an example of no distinguished “heavy” phases and “light” phases.

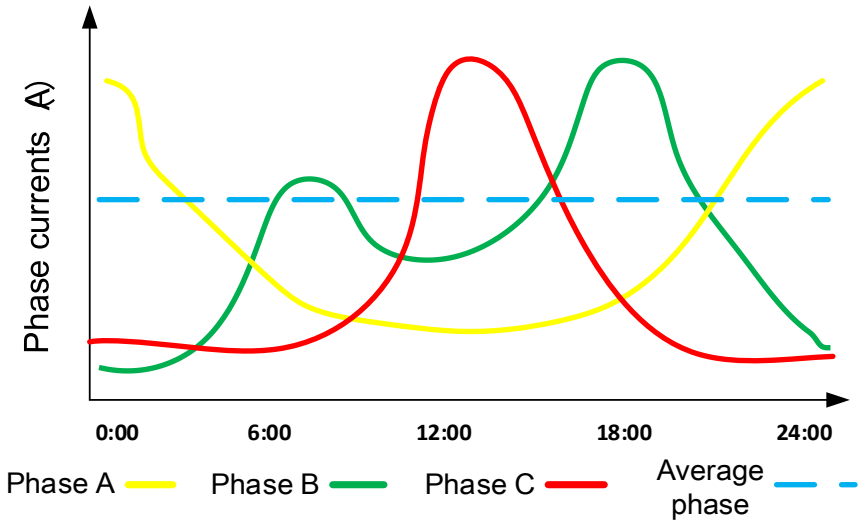


Fig. 2-9 An example of no distinguished “heavy” phases and “light” phases

In Fig. 2-9, the peak time for phases A, B, and C does not occur simultaneously. Shiting peak loads (loads exceed the dashed line) on Phase A, B and C to their off-peak period (when loads below the dashed line) rebalances the three phases.

According to the principle of shifting single-phase loads for phase balancing, reference [25] developed a game-based control to shift single-phase EV charging on time horizons for phase balancing. Reference [7] shifted battery storage systems’ charging and discharging period for phase balancing.

While, using existing LCT devices, such as energy storage systems, EV chargers, and

household PV systems, can theoretically address phase imbalance. there has no business approach that could motivate these customers to provide their LCT devices for phase balancing, especially guiding customers to prioritise predominant imbalance-induced consequences.

2.5. Chapter summary

This chapter does a comprehensive review of existing studies of phase imbalance assessment and phase balancing solutions, including the state of art investigations and industry-adopted approaches.

However, as the thesis has described in Chapter 1.1.2, in reality, the data-scarcity problem invalids all existing literature on phase imbalance assessments and phase balancing decision making. It, therefore, raises two research gaps:

- In making phase balancing investment decisions for individual LV networks, how to assess the imbalance-induced consequences, thus turning phase balancing cost-benefits analysis into possibility.

- Given that off-line phase balancing is the priority choice for the industry [2]. How to make phase swapping guidance before industry field works without the requirement of both customer-side smart meter data and network-side year-round time-series data, thus significantly improving phase swapping's practically in the industrial implementation.

To bridge the above two gaps, this thesis, for the first time, makes three statistical approaches, described in Chapter 3, Chapter 4 and Chapter 5 respectively. These three statisticall appraoches turn massive industrial application of the decision making for phase balancing investments and phase swapping planning into reality.

Chapter 3.

Estimating imbalance-induced energy losses on the three phases for data-scarce LV networks

Chapter contents:

This chapter develops two original methodologies to assessing imbalance-induced phase energy losses and imbalance-induced residual energy losses, respectively, for data-scarce LV networks.

3.1. Chapter summary

Unbalanced three-phase current raise energy losses on distribution wires. For example, the three-phase currents are [3A, 4A, 5A]. The power loss is 50*r,* where *r i*s the equivalent resistance. If the three phases are balanced, where each phase has a current of 4A, its power loss is 48r. In this case, phase imbalance raises energy losses by 4%.

Considering the data-scarcity problem for LV networks, there rises an unsolved question in the industry: how to use available data from data-scarce LV networks (LV networks that do not record time-series phase current data) to assess imbalance-induced phase energy losses. This chapter develops a regression-based approach to get around the unsolved question and assess imbalance-induced phase energy losses for data-scarce LV networks. This approach acquires knowledge from a small sample set of data-rich networks. The acquired knowledge then is applied to data-scarce LV networks to estimate imbalance-induced phase energy losses. In detail, first, this approach extracts features that exist in both data-rich and data-scarce LV networks. Second, the feature vector is transformed by kernel transformation to achieve better regression results [62]. Third, the robust-linear regression model [63] is applied to learn the relationship between the transformed feature and calculated imbalance-induced phase energy loss for data-rich LV networks. Lastly, the trained regression model estimate the imbalance-induced phase energy loss for data-scarce LV networks. Compared to classic regression methods, such as ordinary linear regression [64], tree, SVR [62] and Gaussian process [65], my approach delivers the highest estimation accuracy. Generally, this approach derives over 80% accuracy for 90% of the data-scarce LV networks. This approach is validated by 10-folds cross-validation [66], using data from 800 data-rich LV networks sampled in WPD’s business area.

Chapter 4.2 is cited from the author's published article in IET Generation, Transmission, and Distribution [67]. The structure of this chapter is organised in an alternative-based format, where the indices, equations, tables, figures and titles are numbered independently.

Statement of Authorship

<table>
<tr><td colspan="4">This declaration concerns the article entitled:</td></tr>
<tr><td colspan="4">Assessment of additional phase energy losses caused by phase imbalance for data-scarce LV networks</td></tr>
<tr><td colspan="4">Publication status (tick one)
Draft manuscript [ ] Submitted [ ] In review [ ] Accepted [ ] Published [√]</td></tr>
<tr><td>Publication details (reference)</td><td colspan="3">Fang, L. and Ma, K. (2020), Assessment of additional phase energy losses caused by phase imbalance for data-scarce LV networks. IET Gener. Transm. Distrib., 14: 675-681. https://doi.org/10.1049/iet-gtd.2019.1036</td></tr>
<tr><td colspan="4">Copyright status (tick the appropriate statement)
I hold the copyright for this material [ ] Copyright is retained by the publisher, but I have been given permission to replicate the material here [√]</td></tr>
<tr><td>Candidate's contribution to the paper (provide details, and also indicate as a percentage)</td><td colspan="3">The candidate contributed to / considerably contributed to / predominantly executed the…
Formulation of ideas:
● 80%
● Defining the real problem in assessing imbalance-induced phase energy losses and the solution's implementation challenges, guided by Dr Kang Ma.
Design of methodology:
● 100%
● Customising a robust-linear based regression that learns the relationship between imbalance-induced phase energy losses and the features of data-scarce LV networks.
Experimental work:
● 100%
Presentation of data in journal format:
● 80%
● Organising and writing this article, revised by Dr Kang Ma</td></tr>
<tr><td>Statement from Candidate</td><td colspan="3">This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.</td></tr>
<tr><td>Signed</td><td>Lurui Fang</td><td>Date</td><td>30/07/2021</td></tr>
</table>

Assessment of additional phase energy losses caused by phase imbalance for data-scarce LV networks

Lurui Fang [1], Kang Ma[*] [1]

[1] Department of Electronic and Electrical Engineering, University of Bath, Bath, UK

[*]K.Ma@Bath.ac.uk

Abstract: Unbalanced phase currents, which flow in transformer windings and distribution wires, cause a significant increase (approximately 33%) of phase energy losses in low voltage (LV, 415V) networks. However, these additional phase energy losses (APEL) are hard to calculate for most LV networks. A key challenge is that these LV networks are data-scarce, with only yearly average and maximum phase currents. To estimate the APEL for data-scarce LV networks, this paper proposes a statistical approach that effectively overcomes the above challenge. Firstly, the approach calculates APEL for a sample set of data-rich networks with year-round time-series phase current data. Secondly, features are extracted from these networks by considering: 1) whether the features are strongly correlated to additional phase energy losses; and 2) whether the features can be derived from available data (e.g. yearly average and maximum phase currents) from data-scarce networks. Thirdly, to approximate mappings from the features (derived in stage 2) to the APEL (derived in stage 1), a kernel-based regression model is developed, using the customised features. Given any data-scarce network, its APEL is then estimated by applying the regression model. Cross-validation shows that the statistical approach incurs an average error of 13% for 90% of the data-scarce LV networks, excluding the networks with very low APEL values.

Index terms: phase imbalance, energy losses, regression, feature selection

3.2. Introduction

Phase imbalance is a widespread problem in low voltage (415V, LV) networks in the UK and other countries [2], [68], [32]. According to the data from Western Power Distribution (WPD, a UK distribution network operator), more than 50% of the LV networks suffer from a notable degree of phase imbalance. It is common that the current on the heaviest phase is greater than that on the lightest phase by more than 50% [1] . It should be noted that even if a network has perfectly balanced three phases, there is still an I^2R loss on the phase conductors because of conductor impedance. However, if the three phases are unbalanced, the I^2R loss on the phase conductors would be greater than if the three phases were balanced. The difference is the additional phase energy loss (APEL).

These unbalanced phase currents cause a significant increase in energy losses on the three phases of LV networks: 1) on distribution lines, APELs account for up to 33% of wire energy losses [1]; and 2) in distribution transformers, APELs account for up to 27% energy losses of transformer copper losses [1]. However, APELs are hard to calculate for most LV networks. A key challenge arises from the APEL estimation: a lack of time-series phase current data for the majority of LV networks that are data-scarce. These networks only have yearly average and maximum phase current data collected once a year.

One solution to address the data scarcity challenge is to deploy monitoring devices for more than 900,000 LV networks in the UK. However, this causes a substantial cost. With sufficient data collected, a number of references assess energy losses caused by phase imbalance. Reference [40] assesses the additional copper losses caused by

imbalanced loading for LV transformers. Reference [8] evaluates energy losses in distribution networks with imbalanced three phases. The APELs are calculated for networks with full data. Reference [69] develops network component models (includes load, line, and transformers) to calculate the energy losses for distribution transformers and lines. The APELs are then derived from this model. Reference [68] develops new phase imbalance indices, which are then used to estimate energy losses. Reference [41] uses complex unbalance factors to evaluate the APEL.

Reference [70] develops a combination of clustering and classification approaches to estimate the imbalance-induced energy losses for data-scarce networks. However, Reference [70] focuses on the energy losses in the neutral and ground, caused by phase residual currents. Such energy losses have a fundamentally different mechanism from the APEL, which occurs on the phases. Because of the different mechanisms, this paper uses a completely different methodology from that in [70]. Compared to Reference [70], which uses a combination of clustering and classification, this paper develops a straightforward regression model using customised features. This model achieves a greater estimation accuracy than the approach in Reference [70].

In addition, it is popular to use the load loss factor to estimate the energy losses on each phase [44]. The APEL can be directly calculated if the energy losses on each of the three phases were available. However, the load loss factor k is suggested to be updated every month [44]. This incurs a prohibitively high cost to collect these data every month for the mass population of LV networks throughout the UK. Reference [71] models the correlation between the increase of energy losses and imbalance degrees based on three scenarios, e.g. 1) one phase is overloaded, and the other two phases have light loads; 2) two phases are overloaded, and the other phase has a light load; and 3) the three phases are overloaded, moderately loaded, and lightly loaded,

respectively. Reference [72] develops a statistical approach to estimate energy losses in distribution components (e.g. distribution lines, transformer, etc.) based on load curves. However, Reference [72] does not assess the APEL caused by phase imbalance.

Based on the literature review, a research question arises: to assess the APEL caused by phase imbalance for data-scarce LV networks. This paper makes an original contribution by answering the above research question for the first time. To this end, this paper develops a new customised statistical approach, using customised features, to assess the APELs for data-scarce networks. This approach learns the knowledge from a sample set of 800 data-rich networks (with time-series phase current data throughout a year), then infers the APELs by extrapolating the knowledge to these data-scarce networks.

The customised methodology is designed to be highly practical for distribution network operators (DNOs), who can directly apply the methodology to their business areas. Furthermore, the APEL is one of the key inputs for the cost-benefit analysis of phase rebalancing for data-scarce networks. In addition, it can help the DNOs to assess the additional heating caused by phase imbalance for data-scarce LV networks. This additional heating is one of the key components in analysing the thermal ratings of electric apparatuses (e.g. distribution transformers and lines) in data-scarce LV networks.

The rest of this paper is organised as follows: Chapter 3.2.2 presents the methodology. Chapter 3.2.3 performs case studies. Chapter 3.2.3 discusses the limitation of using load flow analysis for estimating imbalance-induced energy losses. Chapter 3.2.5 concludes this paper.

3.3. Methodology

The statistical approach consists of three stages. Firstly, it calculates the APELs for 800 data-rich networks with time-series phase current data throughout a year. Then, features are selected by considering: 1) whether the features are strongly correlated to the APELs; and 2) whether the features can be obtained from data-scarce networks that only have yearly average and maximum phase currents. Thirdly, a regression model is developed to map the features (derived in Stage 2) to the APELs (derived in Stage 1). Given any data-scarce network that has the feature vector as the input, the APEL is estimated by applying the developed regression model.

The flowchart of the proposed approach is shown as follows:

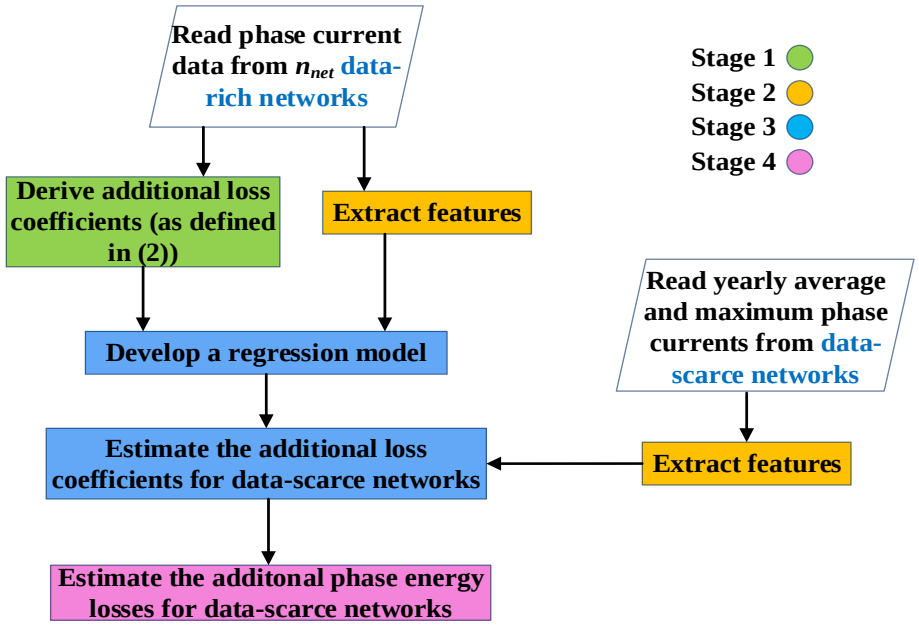


Fig. 3-1 Overview of the statistical approach

The project “Low Voltage Network Template” provides time-series phase current data throughout a year from n_{net} ($n_{net} = 800$) data-rich networks. These networks cover: 1)

a good mixture of urban, suburban and rural areas; and 2) a good mixture of household, commercial and industry loads [1].

3.3.1.Data processing

For data-rich networks, a virtual current is defined as:

$$I_v(t) = \sqrt{I_a(t)^2 + I_b(t)^2 + I_c(t)^2 - 3\left(\frac{I_a(t) + I_b(t) + I_c(t)}{3}\right)^2} \quad \text{3-(1)}$$

where $I_a(t), I_b(t)$, and $I_c(t)$ denote the currents on phase a, b and c, respectively, at time t.

Then an additional loss coefficient is defined as:

$$L_{ac} = \frac{1}{n_y} \sum_{t=1}^{n_y} I_v(t)^2 \quad \text{3-(2)}$$

where I_v is defined in 3-(1); n_y is the length of time-series phase current data throughout a year. The reason for defining this coefficient is to normalise the sum of $I_v(t)^2$ for all data-rich networks. This prevents large values of the sums of $I_v(t)^2$ from causing large root-mean-squared errors, thus improving the accuracy of the regression model.

For most LV networks, their topologies are unknown for the DNO. According to reference [73], loads are assumed to be distributed in a rectangular fashion along the LV networks. This results in the equivalent distribution line resistance being discounted to only 1/3 of the original line resistance, but the transformer resistance is unaffected. Therefore, the APEL is given by [73]:

$$E_{al} = T \cdot L_{ac} \cdot (\tfrac{1}{3} R_D + R_T) \quad \text{3-(3)}$$

where T ($T = 8760$) is the number of hours throughout a year; R_D is the resistance of the distribution line; R_T is the resistance of the transformer winding referred to the LV side.

The resistance values of distribution lines and transformers vary in different LV networks. The key output of this stage is the additional loss coefficient L_{ac}, which will be used for regression later.

3.3.2.Feature extraction

To select the features, two factors are considered: 1) whether the features are strongly correlated to additional loss coefficients (derived in Chapter 3.2.2.1); and 2) whether the features can be derived from the available data (i.e. yearly average and maximum phase currents) from data-scarce networks. Based on the above principles, four features are selected: hypothetical virtual current, maximum current, hypothetical degree of phase imbalance, and root-mean-square of unbalance ratio.

1) The hypothetical virtual current is given by:

$$I_{hv} = \sqrt{{I_{ya}}^2 + {I_{yb}}^2 + {I_{yc}}^2 - 3(\frac{I_{ya} + I_{yb} + I_{yc}}{3})^2} \quad \text{3-(4)}$$

where I_{ya}, I_{yb}, I_{yc} denotes the yearly average phase currents on phases a, b and c, respectively.

2) The maximum current is given by:

$$I_m = \max\left\{I_{yma}, I_{ymb}, I_{ymc}\right\} \quad \text{3-(5)}$$

where I_{yma}, I_{ymb} and I_{ymc} denote the yearly maximum currents on phases a, b and c, respectively; $\max\{\dots\}$ indicates the maximum value of $\{\dots\}$.

3) The hypothetical degree of phase imbalance is given by:

$$DIB_v = \frac{(\max\{I_{ya}, I_{yb}, I_{yc}\} - \frac{I_{ya} + I_{yb} + I_{yc}}{3})}{I_{ya} + I_{yb} + I_{yc}} \quad 3\text{-}(6)$$

where I_{ya}, I_{yb} and I_{yc} are defined in 3-(4).

4) The root-mean-square (*RMS*) of unbalance ratio

Before deriving this *RMS* value, the positive, negative and zero sequence currents are given by:

$$\begin{bmatrix} \dot{I}_1 \\ \dot{I}_2 \\ \dot{I}_0 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & q & q^2 \\ 1 & q^2 & q \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} \dot{I_{ya}} \\ \dot{I_{yb}} \\ \dot{I_{yc}} \end{bmatrix} \quad 3\text{-}(7)$$

where q is $e^{j2\pi/3}$; $\dot{I_{ya}}, \dot{I_{yb}}, \dot{I_{yc}}$ are the yearly average complex current values on phases a, b and c, respectively; the upper dot indicates that these values are complex values, which are 120 degrees apart from each other.

RMS is then given by [74]:

$$RMS = \sqrt{\left|\dot{I}_0\right|^2 + \left|\dot{I}_2\right|^2} \Big/ \left|\dot{I}_1\right| \quad 3\text{-}(8)$$

where $\left|\dot{I}_1\right|$, $\left|\dot{I}_2\right|$ and $\left|\dot{I}_0\right|$ are the magnitudes of $\dot{I}_1, \dot{I}_2$ and $\dot{I}_0$, respectively. A feature vector consisting of the above features is given by:

$$f_v = [I_{hv}, I_{hm}, DIB_h, RMS] \quad 3\text{-}(9)$$

where I_{hv}, I_{hm}, DIB_h and RMS are defined in 3-(4), 3-(5), 3-(6), and 3-(8), respectively.

Through the case study, a high regression accuracy is achieved when considering all the above features. This shows a strong correlation between the selected features and the additional loss coefficients.

3.3.3.Develop the regression model

In this stage, a kernel-based robust linear regression model is developed. It approximates the mappings from the features (derived in Chapter 3.2.2.2) to the additional loss coefficients (derived in Chapter 3.2.2.1) through training on the sample set of the data-rich networks. Then the developed mapping is applied to any data-scarce LV network with the feature vector only to estimate its additional loss coefficient. This value is then converted to the APEL for the data-scarce LV network by applying 3-(3). The reasons for using the kernel-based robust linear regression model are: 1) robust linear regression is a classic regression approach [75]; 2) it is less sensitive to outliers [75]; and 3) the approach allows for a higher regression accuracy compared to alternative classical regression approaches. The comparison will be demonstrated in case studies.

In the first step, a quadratic kernel transformation is used to transform the feature vector from its original space to a vector in a high dimensional Hilbert space [62]. This is because the mapping in the original space is non-linear; the quadratic kernel transformation enables a nearly linear mapping in the high dimensional space. Through such a transformation, the regression accuracy is improved by 43% compared to the ordinary robust linear regression. The quadratic kernel transformation is given by:

$$f_{kv} = [k(f_{v,1}, f_{v,1}), \cdots k(f_{v,i}, f_{v,j}) \cdots, k(f_{v,4}, f_{v,4})] \quad \text{3-(10)}$$

where $k(f_{v,i}, f_{v,j}) = ({f_{v,i}}^{\mathrm{T}} \cdot f_{v,j} + c)^2$;

$f_{v,i}$ and $f_{v,j}$ are the i_{th} and j_{th} variables in the feature vector f_v (as defined in 3-(9)), respectively; c denotes a constant value. Based on this transformation, the feature vector f_v is transformed into a high dimensional kernel feature vector f_{kv}. In this study, f_{kv} is a vector with 16 variables.

Then, a robust linear regression model is developed to approximate the mapping from the kernel feature vector f_{kv} (defined in 3-(10)) to the additional loss coefficients L_{ac} (given by 3-(2)) for data-rich networks, as given by:

$$\begin{bmatrix} L_{ac,1} \\ \vdots \\ L_{ac,n_{net}} \end{bmatrix} = \begin{bmatrix} f_{kv,1} \\ \vdots \\ f_{kv,n_{net}} \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_{net}} \end{bmatrix} \quad \text{3-(11)}$$

where $L_{ac,i}$ is the additional loss coefficients for the i_{th} data-rich LV network, as defined in 3-(2); $f_{kv,i}$ is the kernel feature vector with n_f ($n_f = 16$) columns for the i_{th} data-rich LV network; β is a coefficient vector with n_f rows; ε_i is the regression error for the i_{th} data-rich LV network; $n_{net}(n_{net} = 800)$ is the number of data-rich networks.

To obtain β and ε, an iterative algorithm is presented as follows:

1) Set *i* = 0. The ordinary linear regression [64] is used to derive coefficient vector $\beta^{(i)}$ and error vector $\varepsilon^{(i)}$.

2) According to the derived error vector $\varepsilon^{(i)}$, weighting vector w_{i+1} are given to the training samples (data-rich networks), as high weights are given to samples with low errors. This weight function is defined by:

$$w_{i+1} = \frac{1}{\varepsilon^{(i)}} \quad \text{3-(12)}$$

3) Set $i \to i + 1$. A weighted least square model is used to minimise:

$$\min \sum w_i \, {\varepsilon^{(i)}}^2 \quad \text{3-(13)}$$

After finding all w_i, $\beta^{(i)}$ is given by:

$$\beta^{(i)} = \left({f_{kv}}^T W \, f_{kv} \right)^{-1} {f_{kv}}^T W L_{eo} \quad \text{3-(14)}$$

where $L_{ac}$and f_{kv} are defined in 3-(11); W is the diagonal matrix of individual weights in w_i. Correspondingly, a new $\varepsilon^{(i)}$ is derived in this step.

4) Steps 2) and 3) are repeated until the coefficient vector $\beta^{(i)}$ stabilized.

Detailed implementations of steps 1) – 4) are presented in [76]. After finding β, the additional loss coefficient L_{acs} for any data-scarce LV network is given by:

$$L_{acs} = f_{ksv}\beta \quad \text{3-(15)}$$

where L_{acs} is a scalar. f_{ksv} is the kernel feature vector of the data-scarce network. It has n_f columns. f_{ksv} is given by 3-(10), where f_{ksv} replaces f_{kv}. β is given by 3-(14). β is a vector with n_f rows.

3.3.4. Validation

In this paper, the *k*-fold cross-validation [66] is used to validate my developed approach and derive the estimation accuracy. The reasons for using *k*-fold cross-validation are: 1) the cross-validation avoids using the same data to both develop and validate the developed model; and 2) it ensures a satisfactory tradeoff between bias and variance. In each iteration of the cross-validation, a portion of the data-rich networks are reserved in the validation set and are treated as if they were data-scarce. Their APEL results are estimated by applying my approach, which is trained using the rest of the data-rich networks. However, because the networks in the validation set are indeed data-rich networks, their accurate APEL results can be calculated. This allows for the comparison of the estimated APEL results against the accurate APEL results for validation.

The k-fold cross-validation is detailed as follows. Firstly, the additional loss coefficient L_{ac} (as defined in Chapter 3.2.2.1) are derived as the accurate values for the 800 data-rich networks. The rest steps are described in Fig. 3-2.

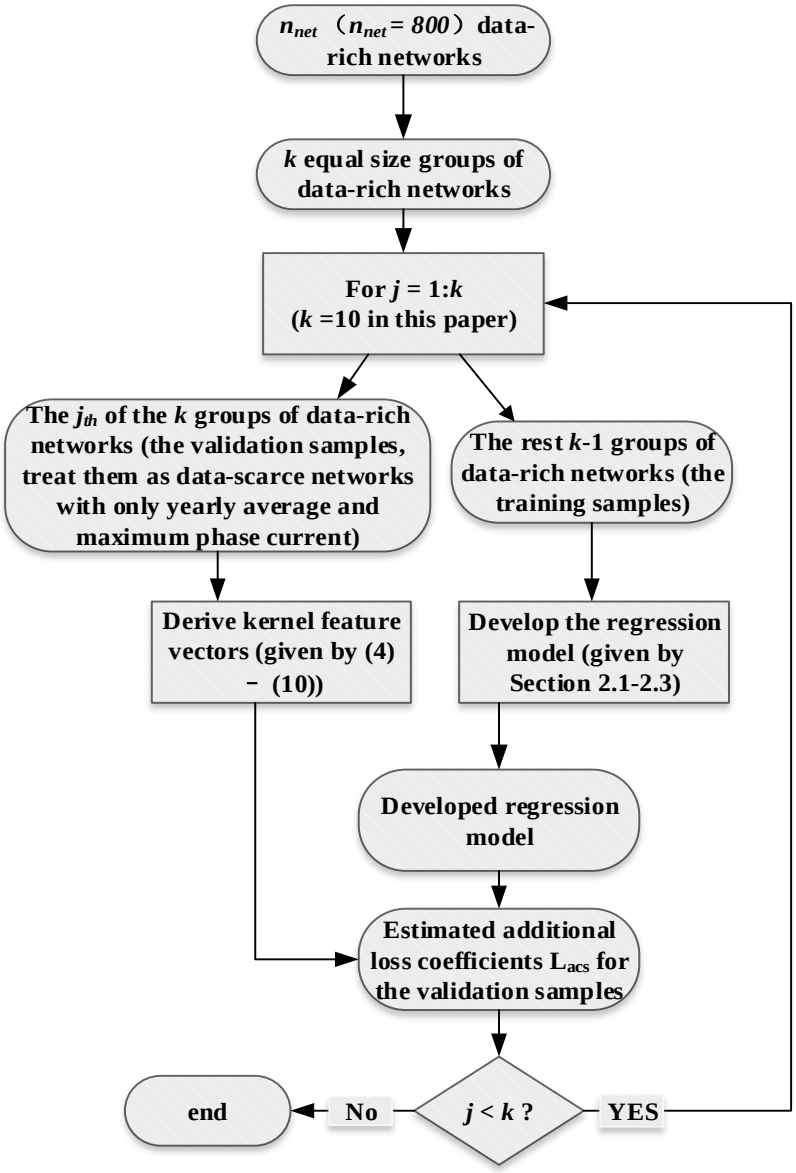


Fig. 3-2 Flowchart of k-fold cross-validation

This paper uses the root-mean-square error (RMSE) to measure the regression performance. The regression performance indicates errors between the accurate

values L_{ac} derived in Chapter 3.2.2.1 and the estimated values L_{acs} derived by applying the *k*-fold cross-validation to the validation samples (treat them as data-scarce networks). This error is given by:

$$e_{rmse} = \sqrt{\frac{\sum_{i}^{n_{net}}\left(L_{ac,i} - L_{acs,i}\right)^2}{n_{net}}} \quad \text{3-(16)}$$

where $n_{net}(n_{net} = 800)$ is the number of validation samples. $L_{acs,i}$ is the estimated additional phase energy loss for the i_{th} validation sample (treat it as if it were a data-scarce network with only yearly average and maximum phase currents); $L_{ac,i}$ is the accurate value (derived in Chapter 3.2.2.1) of additional phase energy loss for the i_{th}validation sample. A lower e_{rmse} indicates a better performance of the developed regression model.

3.3.5. Additional phase energy losses estimation for data-scarce networks

After deriving the additional loss coefficients for data-scarce networks, the APELs are estimated in two scenarios: 1) the resistances of distribution lines are available; and 2) the resistances of distribution lines are unknown.

Given a data-scarce network, its APEL is given by 3-(3), where L_{acs} replaces L_{ac}. L_{acs} is given by 3-(15).

For scenario 1), the APELs are directly calculated by applying 3-(3). For scenario 2), the APELs are calculated using typical wire resistances for urban, suburban and rural networks in the UK. The typical wire resistances for urban, suburban and rural networks are 0.064Ω, 0.282Ω and 0.32Ω, respectively [66].

3.4. Case study

This Chapter presents numerical results: 1) Chapter 3.2.3.1 gives the additional loss coefficients and corresponding features for the 800 data-rich networks; 2) Chapter 3.2.3.2 presents the regression results; 3) Chapter 3.2.3.3 presents the APEL results for data-scarce networks; and 4) a discussion is given in Chapter 3.2.3.4.

3.4.1. Data processing and feature extraction

In this Chapter, for the 800 data-rich LV networks, the APEL are firstly derived and presented in Fig. 3-3.

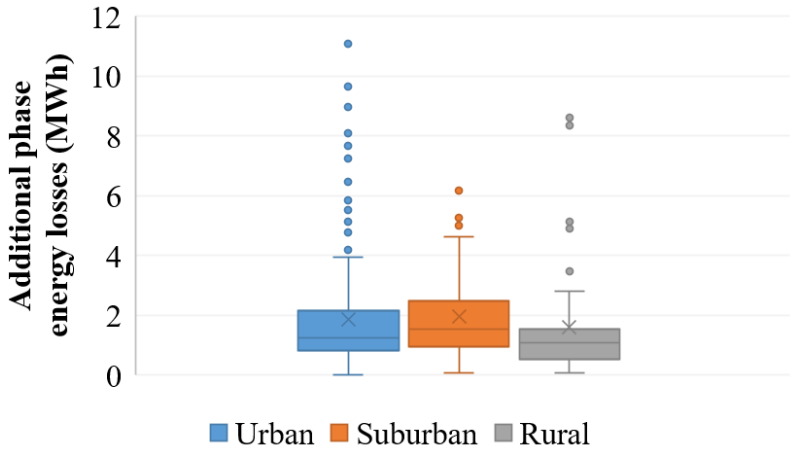


Fig. 3-3 The additional phase energy losses for data-rich networks in urban, suburban and rural areas.

Fig. 3-3 is the range of the APELs (shown in box plot) for the 800 data-rich networks. For example, the blue dot indicates the outliers. The upper and bottom blue lines indicate the maximum and minimum APELs for urban LV networks. The line in the blue box is the average APEL for urban LV networks. The blue box indicates the range of APELs for most urban LV networks. In Fig. 3-3, the average APELs are 1.79 MWh, 1.95 MWh and 1.59 MWh for LV networks in urban, suburban and rural areas,

respectively. For rural LV networks, the average and maximum APEL account for 0.21% and 1.21%, respectively, of the yearly distributed energy. For suburban LV networks, the average and maximum APEL account for 0.44% and 1.42%, respectively, of the yearly distributed energy. For rural LV networks, the average and maximum APEL account for 0.68% and 3.66%, respectively, of the yearly distributed energy. Furthermore, for LV networks in suburban and rural areas, the APEL account for up to: 1) 33% of the total wire energy losses; and 2) 27% of the total transformer copper losses.

Then, to develop the regression model, the additional losses coefficients L_{ac} and corresponding features (e.g. hypothetical virtual current I_{hv}, hypothetical maximum current I_{hm}, Hypothetical degree of phase imbalance DIB_v, Root mean squares of unbalance ratio RMS) are derived. Example are given as follows:

TABLE 3-1 EXAMPLES OF THE ADDITIONAL PHASE ENERGY LOSSES COEFFICIENTS AND CORRESPONDING FEATURES FOR DATA-RICH NETWORKS

	$\mathrm{L_{ac}}$	$\mathrm{I_{hv}}$	$\mathrm{I_{hm}}$	$\mathrm{DIB_v}$	RMS
1	1637	8.93	413.4	0.01	0.06
2	5799	53.8	614.8	0.05	0.3
3	6492	51.9	1235.3	0.02	0.11
4	2801	32.5	508.6	0.03	0.16
5	836	8.95	330.5	0.02	0.06

Thirdly, the regression error (shown in root-mean-squared error (RMSE) and mean-average-percentage error (MAPE)) from the kernel-based robust regression is used to validate the choice of these features. A lower regression error indicates a better selection of features. This validation is performed for four scenarios: 1) only $\mathrm{I_{hv}}$ is used as the feature to develop regression models; 2) $\mathrm{I_{hv}}$ and $\mathrm{I_{hm}}$ are used as the features to develop regression models; 3) $\mathrm{I_{hv}}$, $\mathrm{I_{hm}}$and $\mathrm{DIB_v}$ are used as the features to develop regression models; 4) excluding $\mathrm{L_{ac}}$, all four features in TABLE 3-1 are used as the

features to develop regression models. The validation results are presented in TABLE 3-2.

TABLE 3-2 REGRESSION ERROR IN THE ABOVE SCENARIOS

Scenario	1)	2)	3)	4)
RMSE	1163	961	715	632
MAPE	41.9%	33.4%	22.3%	19.7%

In TABLE 3-2, the RMSE and MAPE decrease with an increasing number of features used for regression. The results justify the choice of all the customised features in this paper.

3.4.2.Regression results

In this Chapter, a kernel-based robust linear regression model is developed. The regression accuracy is significantly higher than that of ordinary robust linear regression. Through k-folds validation (defined in Chapter 3.2.2.4), the validation results are shown in Fig. 3-4 and Fig. 3-5.

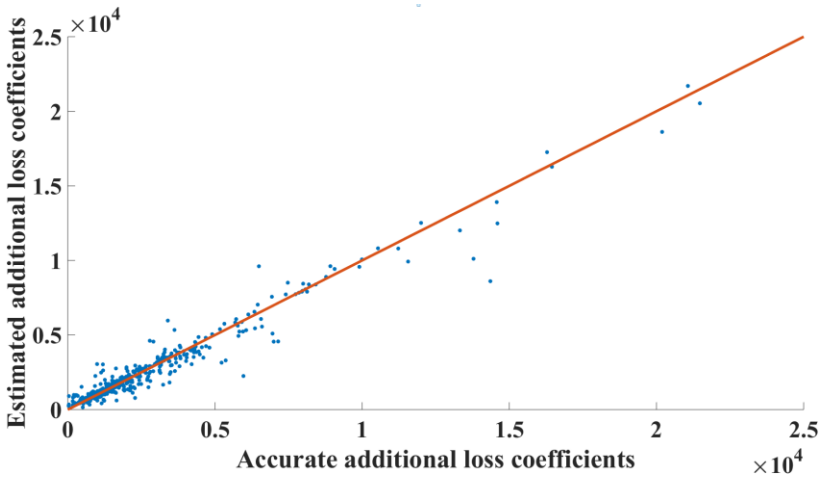


Fig. 3-4 The validation results of kernel-based robust linear regression

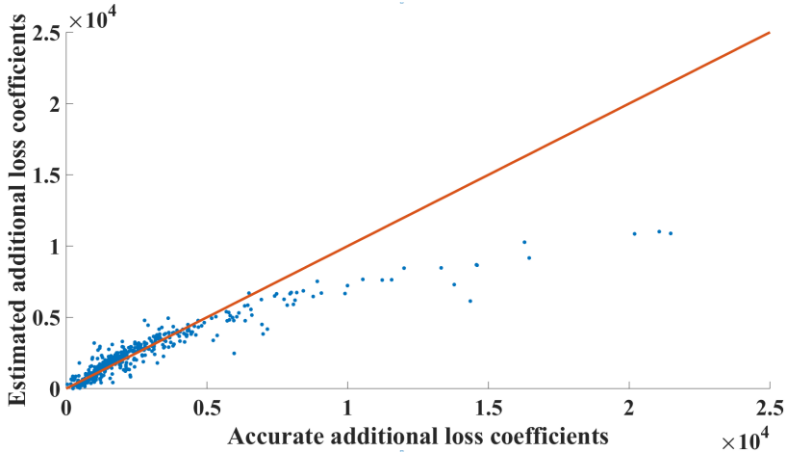


Fig. 3-5 The validation results of ordinary robust linear regression

In Fig. 3-4 and Fig. 3-5, the x-axis represents the accurate additional loss coefficients which are given by Equation 3-(2) for 800 data-rich LV networks. The y-axis represents the estimated additional loss coefficients, when these data-rich LV networks are treated as data-scarce in the *k*-folds validation (shown in Chapter 3.2.2.4). The red line indicates if the additional loss coefficients are perfectly estimated by regression models. If the blue dots are closer to the red line, it indicates a higher regression accuracy. The estimated additional loss coefficients delivered by kernel-based robust linear regression are much closer to the red line than that from ordinary robust linear regression. The root-mean-squared error (RMSE) delivered by kernel-based robust linear regression is 632, which is 43% lower than that by ordinary robust linear regression.

Furthermore, the kernel-based robust regression achieves a higher regression accuracy compared to other classic regression approaches. The comparison is given as follows:

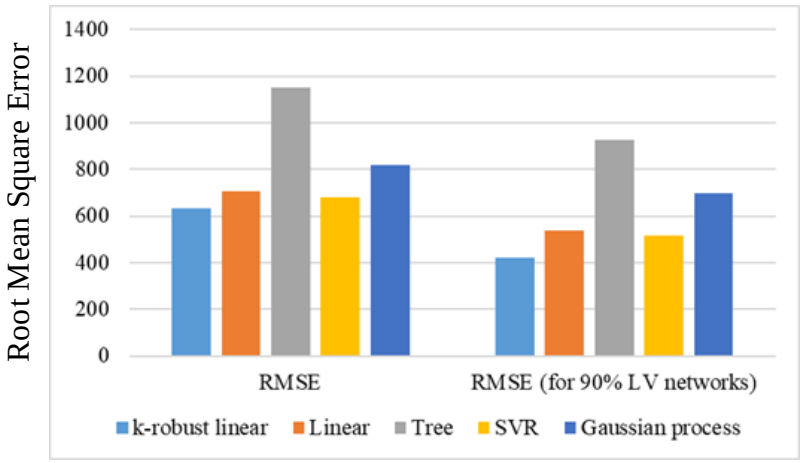


Fig. 3-6 Comparison of the regression approaches

In Fig. 3-6, the kernel-based robust linear regression achieves almost the same RMSE as that by the support vector machine. However, when excluding 10% outliers (which presents lower regression accuracies than most LV networks), the RMSE, delivered by kernel-based robust linear regression, is lower than that from the support vector machine by 12% and other regression approaches by up to 37%. My methodology has an RMSE of slightly above 400, whereas alternative approaches have RMSE values of above 500. The reduction in RMSE is attributed to the robust linear regression, kernel transformation and the customisation of features in my methodology. Further, when excluding 10% outliers, the k-robust linear regression only incurs a MAPE of 13%, i.e., on average, the estimated APEL is only 13% away from its accurate value. This estimation error is acceptable as these data-scarce LV networks only have the yearly average and maximum phase currents. However, linear regression, tree regression, SVR and Gaussian process regression incur greater MAPEs of 17.3%, 32.7%, 16.5% and 23.9%, respectively.

3.4.3. Assessments of additional phase energy losses for data-scarce networks

After developing the regression model and calculating the additional loss coefficients for data-scarce LV networks, the APELs are derived by 3-(3), where L_{acs} replaces L_{ac}. L_{acs} is given by 3-(15). The k-folds validation results are shown as follows:

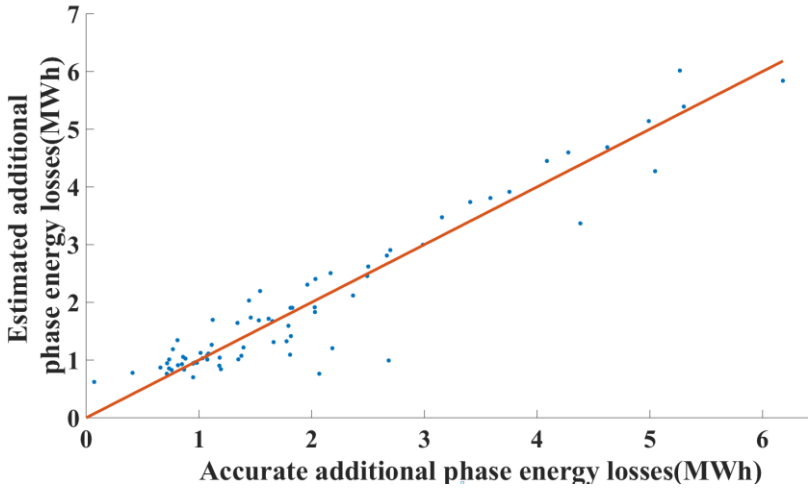


Fig. 3-7 The estimation of additional phase energy losses for LV networks in urban areas

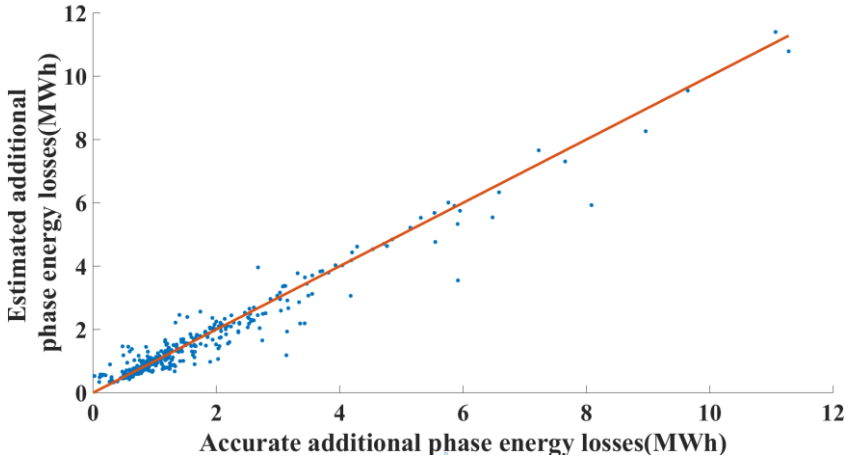


Fig. 3-8 The estimation of additional phase energy losses for LV networks in suburban

areas

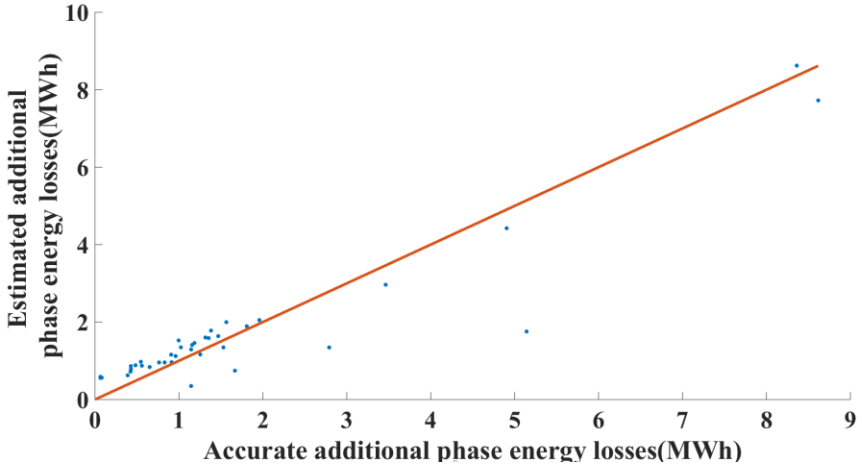


Fig. 3-9 The estimation of additional phase energy losses for LV networks in rural areas

In Fig. 3-7, the estimated average APEL is 1.746 MWh (which costs £314 if the electricity price is £0.18/kWh) for data-scarce urban LV networks. The average estimation error is 19.14% for 90% of the urban networks. In Fig. 3-8, the estimated average APEL is 1.954 MWh (which costs £352 if the electricity price is £0.18/kWh) for data-scarce suburban LV networks. The average estimation error is 11.81% for 90% of the suburban networks. In Fig. 3-9, the estimated average APEL are 1.531MWh (which costs £276 if the electricity price is £0.18/kWh) for data-scarce rural LV networks. The average estimation error is 12.19% for 90% of the data-scarce LV networks in rural areas.

The following figure presents the estimation accuracy of the proposed approach for LV networks with different imbalance degrees, which are defined in [77].

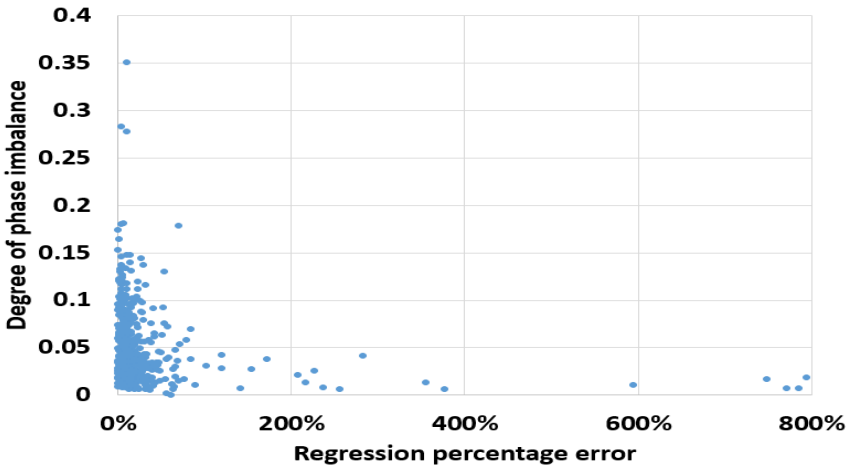


Fig. 3-10 Regression errors for LV networks with different degrees of imbalance

In Fig. 3-10, with the increase of the degree of phase imbalance, my proposed approach delivers lower percentage error, i.e. a higher estimation accuracy is achieved for highly imbalanced LV networks. For LV networks with 0.1 or higher degrees of phase imbalance, the average percentage regression error is 11.7%.

3.4.4.Discussions

In this study, my developed approach delivers about 13% percentage error in estimating the APEL for 90% of the data-scarce LV networks. This error is satisfactory because the developed approach uses minimal data (e.g. yearly average and maximum phase currents, which exists in most LV networks) to assess the year-round APEL for data-scarce networks. A higher regression accuracy can be derived if more input data are used for data-scarce networks. A trade-off is thus required by the DNOs, i.e. the DNOs should decide if it is worth collecting more data for a slightly higher regression accuracy, as more input data means more costs on data collection. In addition, for LV networks in the urban area, the estimation error of APEL is higher than that for LV networks in suburban and rural areas by up to 50%. However, the higher estimation error for urban networks is acceptable. It is because, according to this study,

urban networks correspond to very minimal APEL (only £165.6 which accounts for 9.5% of the APEL for rural networks), which are not the focus for the DNOs. For the critical focus networks (e.g. LV network which presents higher APEL in suburban and rural areas), this study delivers significant lower estimation errors, which are 11.81% and 12.19% for suburban and rural networks, respectively.

To apply this approach in other countries, two points should be considered when choosing the data-rich networks: 1) there should be at least 800 data-rich networks to be collected; and 2) these data-rich LV networks should be representative. They should cover a good mix of geographical areas (urban, suburban, and rural) and customer composition (domestic, commercial, and industrial). A higher estimation accuracy would be achieved if the training data are more representative.

For the DNOs, this paper developed an effective and efficient approach to assess the APEL. For 90% of the data-scarce LV networks (excludes 10% outlier networks), the estimation error is about 13%. In this study, it is appropriate to exclude these 10% outliers. It is because all these outliers have low APEL.

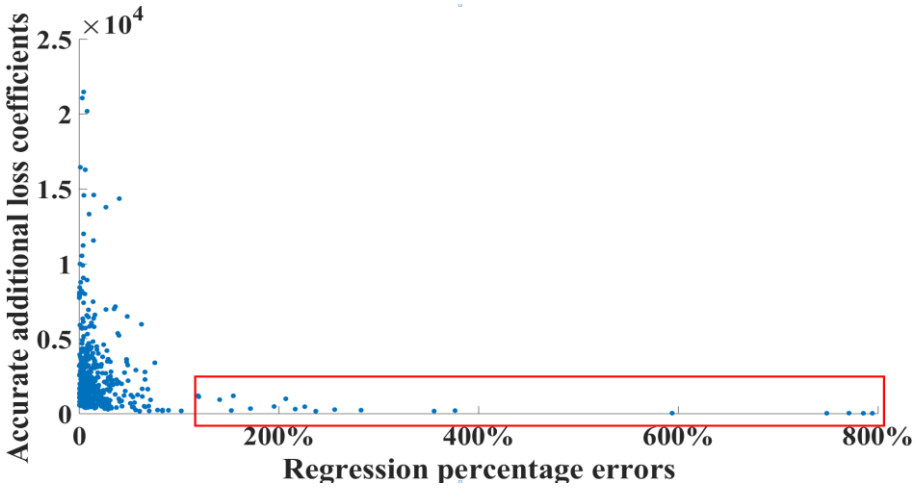


Fig. 3-11 Regression errors for outlier networks

In Fig. 3-11, the outliers are enclosed in the red box. They present significant regression errors by up to 800%. However, these outliers show very low APEL, which are only up to 0.3MWh ((which costs £54 additional losses if the electricity price is £0.18/kWh)). These outliers are thus out of focuses by the DNOs. Furthermore, for LV networks with significantly higher APEL, this study delivers much lower estimation errors. It is, therefore, appropriate to exclude these outliers.

For data-scarce LV networks, the available data of maximum phase currents can be directly obtained from maximum phase current indicators. The yearly average phase currents can be obtained through: 1) the remote telemetry unit (RTU) device on the high voltage side of LV transformers. The data on the high voltage side are then transformed, referred to the low voltage side. 2) The relay protection device if the device has a metering function [66]. 3) The energy meters if they record the data of the three phases separately. In addition, a recent project, OpenLV, sponsored by Western Power Distribution and undertaken by EA Technology, monitors a range of LV (11kV/415V) substations, and the collected data include the average phase current values [66].

It is appropriate to use regression approaches for assessing the additional phase energy losses for data-scarce LV networks. This is because: 1) it is common to use regression approaches to estimate or predict unknowns in both data science [64], [75] and power systems [78], [79]. 2) Through k-fold cross-validation, my approach delivers a satisfactory regression accuracy, where the average percentage error is 13% for 90% of the LV networks.

For LV networks that have high APELs (over 2.5 MWh) throughout a year, the approach delivers an accuracy of 87.3%, which is greater than the accuracy of the methodology from reference [70] by 23.7%. For LV networks less than 2.5 MWh APELs,

this paper and reference [70] deliver similar estimation accuracies.

For comparison, the additional energy losses are also calculated by applying power flow analysis. However, the power flow analysis incurs unacceptably large errors when estimating the APELs for data-scarce LV networks. Given any data-scarce LV network with only yearly average phase currents and no topology, the process for calculating APEL through power flow analysis is detailed as follows: 1) assuming the loads are distributed in a rectangle distribution [73], calculate the energy losses using the unbalanced yearly average phase currents as the input. 2) Calculate the energy losses using the balanced yearly average phase currents as the input. 3) Calculate the APEL, which is the difference between the energy losses obtained in Steps 2) and 3). Through validation, when excluding 10% outliers, the power flow analysis incurs an average MAPE of 237% in the estimation of the APELs for the 800 LV networks. This error is unacceptably large, proving that the power flow analysis is not suitable for the estimation of the APELs for data-scarce LV networks. In contrast, the methodology developed by this paper is suitable for this task, and it incurs the minimum error compared to alternative approaches.

3.5. Estimating imbalance-induced energy losses on the three phases by load flow analysis

The last paragraph in Chapter 3.2.3.4 discusses why traditional power flow analysis is not applicable for addressing imbalance-induced phase energy loss estimation for LV networks. This Chapter is the expanding contents of this discussion.

First, since the majority of LV networks within the UK have no properly documented topology, transformer-side time-series data and customer-side smart metering data, the traditional power flow analysis can only be utilised based on substantial assumptions.

These assumptions and calculation steps are presented as follows:

1) assuming the loads are distributed in a rectangle distribution [73], calculate the energy losses using the unbalanced yearly average phase currents as the input, as given by:

$$E_{ibp} = \sum_{t} \frac{8}{15} \left({I_{ya}}^2 + {I_{yb}}^2 + {I_{yc}}^2 \right) R \quad \text{3-(17)}$$

where I_{ya}, I_{yb}, I_{yc} denote the yearly average phase currents on phases a, b and c, respectively.

2) Calculate the energy losses using the balanced yearly average phase currents as the input, as given by

$$E_{bp} = 3 \sum_{t} \frac{8}{15} \left(\frac{I_{ya} + I_{yb} + I_{yc}}{3} \right)^2 R \quad \text{3-(18)}$$

3) Calculate the APEL, which is the difference between the energy losses obtained in Steps 2) and 3), as given by:

$$\text{APEL} = E_{ibp} - E_{bp} \quad \text{3-(19)}$$

Given the same data from 800 LV networks as those in Chapter 3.2.4, through validation, the estimation results are presented in Fig. 3-12.

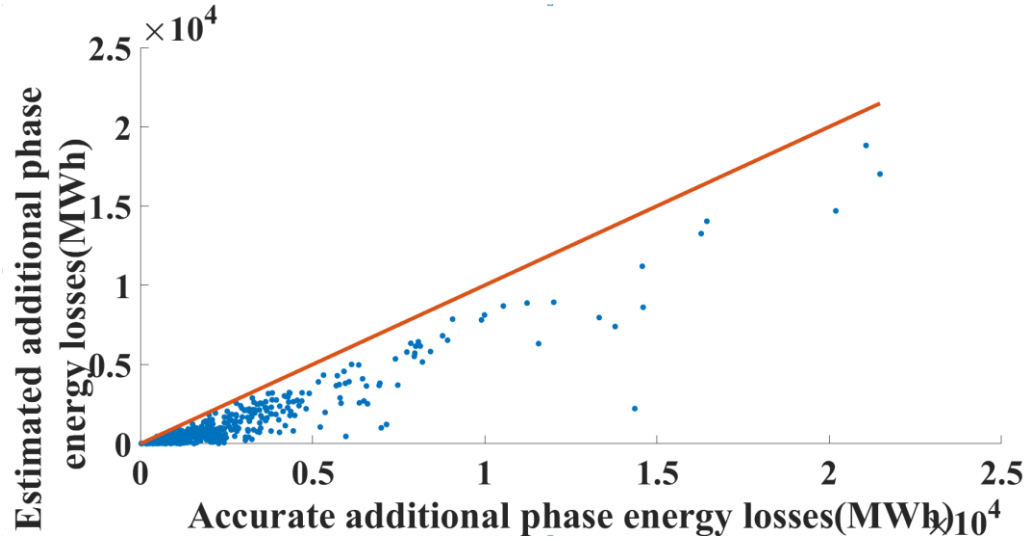


Fig. 3-12 The estimated additional phase energy losses by load flow analysis

In Fig. 3-12, when excluding 10% outliers, the power flow analysis incurs an average MAPE of 237% in the estimation of the APELs for the 800 LV networks. This error is unacceptably large, proving that the power flow analysis is not suitable for the estimation of the APELs for data-scarce LV networks. In contrast, the methodology developed by this paper is suitable for this task, and it incurs the minimum error compared to alternative approaches.

3.6. Conclusions

This study resolves a previously unanswered question: to assess the additional phase energy losses caused by phase imbalance for data-scarce low voltage (415V, LV) networks. To this end, a new statistical approach is developed with customised features. The approach learns the knowledge from 800 data-rich LV networks and then infers the additional phase energy losses for data-scarce LV networks.

Case studies reveal that: for 90% of the data-scarce LV networks in urban, suburban and rural areas, the average regression accuracies are 80.6%, 88.2% and 87.8%,

respectively. These accuracies are satisfactory, as my developed approach uses minimal data (only yearly average and maximum phase currents) to assess the additional phase energy losses.

3.7. Chapter summary

This chapter develops a novel approach to address one unsolved engineering question for data-scarce LV networks: assessing imbalance-induced phase energy losses. Through validation, this approach delivers over 80% estimation accuracy for over 90% data-scarce LV networks. Having accurate estimated imbalance-induced phase energy losses is significant in performing the cost-benefit analysis for phase balancing investments for data-scarce LV networks.

This chapter finds out that directly utilising the existing data from data-scarce LV networks as the input features do not perform satisfactory estimation accuracy. Converting the existing data into other definitions by corresponding engineering equations promotes estimation accuracy. For example, in this chapter, I covert the yearly average phase current and the yearly maximum phase current into the data of hypothetical degree of phase imbalance (defined in equation 3-(6)) and the data of the root-mean-square (*RMS*) of unbalance ratio (defined in equation 3-(8)). These conversions increase the overall estimation accuracy of 14%.

Moreover, this chapter reveals that: for 90% of the data-scarce LV networks in urban, suburban and rural areas, the average regression accuracies are 80.6%, 88.2% and 87.8%, respectively. This indicates that urban networks perform greater data limitation than suburban and rural networks. The scale of the limitation of the data indicates the scale of the natural error in using existing data for realising estimating or forecasting. And this error cannot be addressed by mathematic solutions.

Last, this chapter reveals that outliers exist in this study. In Fig. 3-11, the outliers are enclosed in the red box. They present significant regression errors by up to 800%. However, these outliers show very low APEL, which are only up to 0.3MWh ((which costs £54 additional losses if the electricity price is £0.18/kWh)). Therefore, before training the developed approach, it should remove the LV networks with significant lower APELs from the training data. This prevents the tuned parameters of the trained model from the impact of the outlier LV networks, thus improving the estimation accuracy for the majority of non-outlier LV networks.

Chapter 4.

Estimating imbalance-induced energy losses on the residual path for data-scarce LV networks

Chapter contents:

This chapter develops two original methodologies to assessing imbalance-induced phase energy losses and imbalance-induced residual energy losses, respectively, for data-scarce LV networks.

4.1. Chapter summary

Phase imbalance causes the appearance of phase residual current, which flows from customers to transformers' neutral point via neutral wires or the ground, thus producing energy losses on the phase residual path. The calculation of phase residual current details in Chapter 2.2.1. However, assessing imbalance-induced residual energy losses for a mass-scale of LV networks remains an unsolved question in the industry because of the data-scarcity problem for the majority of LV networks.

This chapter develops a range-estimation-based approach, named clustering, classification and range estimation (CCRE), to address this unsolved problem. This approach has a similar basic principle to the approach detailed in Chapter 3.2. Unlike Chapter 3.2, this chapter uses a range estimation for assessing imbalance-induced residual losses, thus accommodating future phase imbalance changes and being more credible in making long-term phase balancing investment decisions for DNOs. In detail, first, the CCRE approach divides a sample set of data-rich LV networks into groups, using hierarchical clustering [80]. For each group, there exists a probability density function derived from the imbalance-induced residual losses of data-rich networks within this group. Then, MSVM [81] classifies each data-scarce LV network into one of the derived groups with the nearest distance. If a data-scarce LV network is classified into a given group, this group's probability density function of the imbalance-induced residual losses applies to that data-scarce LV network. Lastly, Chebyshev's inequality [82] is applied to narrow down the probability density to the 89% confidence interval. Through 10-folds cross-validation validation, this CCRE approach derives accurate range estimation of imbalance-induced residual losses for over 82% of data-scarce LV networks.

The rest of this chapter is cited from the author's published article in IEEE transactions

on Power systems [70]. The structure of this chapter is organised in an alternative-based format, where the indices, equations, tables, figures and titles are numbered independently.

Statement of Authorship

<table>
<tr><td colspan="4">This declaration concerns the article entitled:</td></tr>
<tr><td colspan="4">A Statistical Approach to Estimate Imbalance- Induced Energy Losses for Data-Scarce Low Voltage Networks</td></tr>
<tr><td colspan="4">Publication status (tick one)
Draft manuscript [ ] Submitted [ ] In review [ ] Accepted [ ] Published [√]</td></tr>
<tr><td>Publication details (reference)</td><td colspan="3">Fang, L., Ma, K., Li, R., Wang, Z., & Shi, H. (2019). A Statistical Approach to Estimate Imbalance-Induced Energy Losses for Data-Scarce Low Voltage Networks. IEEE Transactions on Power Systems, 34(4), 2825-2835. https://doi.org/10.1109/TPWRS.2019.2891963</td></tr>
<tr><td colspan="4">Copyright status (tick the appropriate statement)
I hold the copyright for this material [ ] Copyright is retained by the publisher, but I have been given permission to replicate the material here [√]</td></tr>
<tr><td>Candidate's contribution to the paper (provide details, and also indicate as a percentage)</td><td colspan="3">The candidate contributed to / considerably contributed to / predominantly executed the…
Formulation of ideas:
● 80%
● Defining the real problem in assessing imbalance-induced residual energy losses and the solution's implementation challenges, guided by Dr Kang Ma.
Design of methodology:
● 90%
● Customising a clustering, classification and range-estimation (CCRE) approach that learns the relationship between imbalance-induced phase residual losses and the features of data-scarce LV networks, guided by Dr Kang Ma.
Experimental work:
● 100%
Presentation of data in journal format:
● 80%
● Organising and writing this article, revised by Dr Kang Ma, Dr Ran Li, Dr Zhaoyu Wang, and Dr Heng Shi</td></tr>
<tr><td>Statement from Candidate</td><td colspan="3">This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.</td></tr>
<tr><td>Signed</td><td>Lurui Fang</td><td>Date</td><td>30/07/2021</td></tr>
</table>

A Statistical Approach to Estimate Imbalance- Induced Energy Losses for Data-Scarce Low Voltage Networks

Lurui Fang, Student Member, IEEE, Kang Ma, Member, IEEE, Ran Li, Member, IEEE, and Zhaoyu Wang, Member, IEEE, Heng Shi

Abstract—Phase imbalance in the UK and European low voltage (415V, LV) distribution networks causes additional energy losses. A key barrier against understanding the imbalance-induced energy losses is the absence of high-resolution time-series data for LV networks. It remains a challenge to estimate imbalance-induced energy losses in LV networks that only have the yearly average currents of the three phases. To address this insufficient data challenge, this paper proposes a new customized statistical approach, named as the CCRE (Clustering, Classification, and Range Estimation) approach. It finds a match between the network with only the yearly average phase currents (the data-scarce network) and a cluster of networks with time-series phase current data (data-rich networks). Then CCRE performs a range estimation of the imbalance-induced energy loss for the cluster of data-rich networks that resemble the data-scarce network. The Chebyshev's inequality is applied to narrow down this range, which represents the confidence interval of the imbalance-induced energy loss for the data-scarce network. Case studies reveal that, given such few data from the data-scarce networks, more than 80% of these networks are classified into the correct clusters, and the confidence of the imbalance-induced energy loss estimation is 89%.

Index Terms— energy loss, low voltage, phase imbalance, power distribution, three-phase power

4.2. Introduction

Imbalance-induced energy losses in the UK and European low voltage (415, LV) distribution networks account for up to 35% of the energy losses on distribution wires [7]. This is mainly due to the significant phase imbalance in the UK's LV networks [30], [31], [16]=. Data from Western Power Distribution (a UK distribution network operator) show that over 50% of their LV networks have the peak current of the "heaviest" phase exceeding that of the "lightest" phase by more than 50%, e.g. it is common to have a peak current of 300 A on one phase and 150 A on another phase, causing the phase residual current to be comparable to or even larger than phase currents [1]. The phase residual current then causes an imbalance-induced energy loss. Imbalance-induced energy losses are also widespread in distribution networks in other countries [6], [32]. Therefore, understanding imbalance-induced energy losses are important for distribution network operators (DNOs) to evaluate the total cost of phase imbalance and the potential benefit of phase balancing [77],[83].

There exist a number of references that focus on imbalance-induced energy losses. Reference [13] calculated the energy loss on the neutral wire of overhead lines in the distribution network, using Carson's equations to model the lines. Reference [14] calculatedd neutral energy losses, based on the ratio between the equivalent neutral line resistance and line resistance of a transposed three-phase line. Reference [39] calculated the neutral energy loss caused by non-linear three-phase loads. Reference [8] calculated the neutral energy loss in medium-voltage distribution networks due to load imbalance. Reference [84], [85] calculated the energy losses in distribution networks, including energy losses on both the phases and the neutral wire.

The above references all require networks to have high-resolution time-series data (e.g., data collected every 15 minutes or of a comparable resolution) or load curves.

However, only a small portion of LV networks, the data-rich networks, have high-resolution time-series data, whereas the majority of LV networks only have data collected once a year, i.e., they are data-scarce networks. Therefore, a major challenge to understanding imbalance-induced energy losses is the lack of data. Existing imbalance-induced energy loss estimation methods are not applicable to data-scarce networks.

This paper makes the following original contributions:

1) It for the first time estimates imbalance-induced energy losses for data-scarce networks.

2) To achieve 1), this paper proposes a new customized statistical approach named as CCRE, which consists of three stages: Clustering, Classification, and Range Estimation.

The CCRE approach overcomes the insufficient data challenge by finding a cluster of data-rich networks whose features match the data-scarce network through clustering and classification, using only the yearly average currents of the three phases as the feature. Then this approach performs a range estimation of the imbalance-induced energy loss for the cluster of data-rich networks that resemble the data-scarce network. This range is narrowed down by applying the Chebyshev's inequality formula to counter the impact of outliers. This is the confidence interval of the imbalance-induced energy loss for the data-scarce network.

Because the yearly average phase currents are widely available data in LV networks, this research enables the DNOs to estimate imbalance-induced energy losses on a mass scale across their business area, without the need to deploy high-resolution monitoring devices. This is economically appealing in terms of significant cost savings. According to [86], if all UK's 900,000 LV networks were to be made data-rich, the total

cost of deploying and maintaining pervasive monitoring systems would be approximately two billion British pounds, which can be saved. The proposed method enables the DNO to evaluate a key cost of phase imbalance for the majority of the LV networks that are data-scarce because imbalance-induced energy losses constitute a cost, which occurs year by year until the three phases are rebalanced. This cost is a key input for the cost-benefit analysis of phase balancing solutions.

The rest of this paper is organized as follows: Chapter 4.3 presents the clustering and classification methodology. Chapter 4.4 presents the range estimation of the imbalance-induced energy loss. Chapter 4.5 performs case studies. Chapter 4.6 discusses increasing visible data for data-scarce LV networks for improving estimation accuracy. Chapter 4.7 concludes this paper.

4.3. Methodology

To calculate the imbalance-induced energy loss, two variables, phase residual currents and the impedance data, are required as inputs. However, these two variables are not available in the UK's data-scarce LV networks, which take the majority. For data-scarce networks, the protection systems (e.g. Schneider Sepam series 20) in the substations record the yearly average currents of the three phases [87]. On the other hand, I have time-series phase current data collected from N (in this case, N = 800, but the methodology supports a generic dataset) data-rich LV substations throughout a year at an interval of 15 minutes. These substations, within Western Power Distribution (a UK DNO)'s business area, cover a good mix of geographical areas (urban, suburban, and rural) and customer composition (domestic, commercial, and industrial). For example, Cardiff city centre is selected as an urban area with a large number of commercial customers; Monmouthshire is selected as a representative rural area [1]. These data are the deliverables of the project "Low Voltage Network Templates". Reference [1]

presents a detailed description of these data and this project.

To estimate the phase residual currents for any data-scarce LV network using the available data from the 800 networks, the CCRE approach is proposed. The reason for having the clustering stage is to extract representative characteristics of the phase residual currents (expressed in the form of cumulative density functions) from the 800 data-rich networks, thus transforming the 800 data-rich networks into a few representative classes. Then, the purpose of the classification stage is to find the best match between the data-scarce network and one of the representative classes. Finally, the reason for applying the range estimation is to account for the uncertainty in the imbalance-induced energy loss estimation. Multiple scenarios on the impedance are considered. The overall flowchart of the CCRE approach is presented in Fig. 4-1. It should be noted that all input current data are magnitudes only.

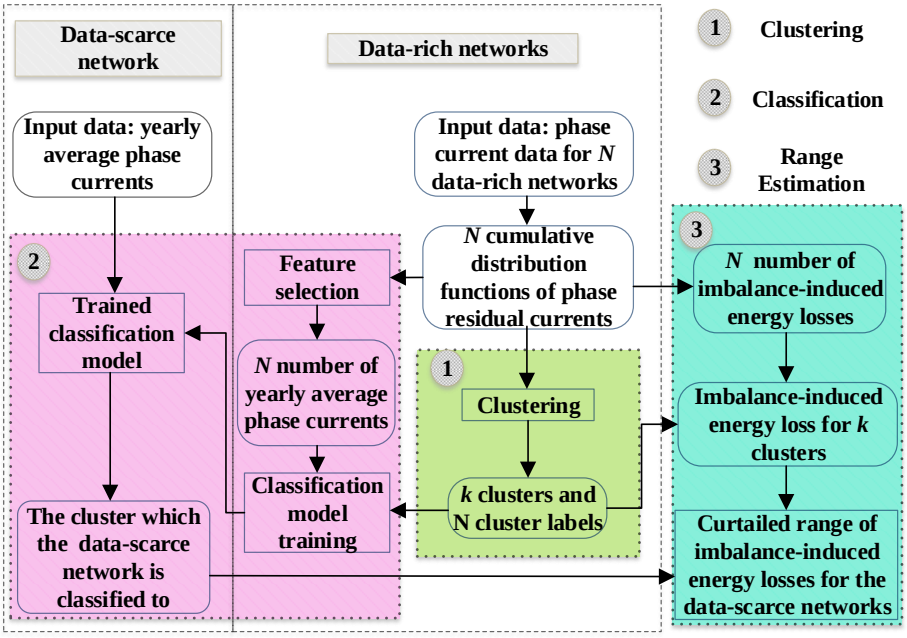


Fig. 4-1 Overview of the CCRE approach

4.3.1.Data pre-processing

The phase residual current $I_{prc}(t)$ is a key variable. For the 800 data-rich LV networks with time-series phase current data, the time-series phase residual current is given by

$$I_{prc}(t) = \left[I_a^2(t) + I_b^2(t) + I_c^2(t) - I_a(t)I_b(t) - I_b(t)I_c(t) - I_a(t)I_c(t)\right]^{1/2} \quad \text{4-(1)}$$

where $I_a(t), I_b(t), I_c(t)$ denote the currents on phases *a*, *b*, and *c* at time *t*, respectively.

In reality, the time-series of phase residual currents for different LV networks have different lengths because there are minor missing data. This paper resolves this problem by transforming each time-series phase residual currents into a cumulative distribution function (CDF). This is suitable because this paper is only concerned about the imbalance-induced energy loss over a year (this is the basis for calculating the annual cost of the imbalance-induced energy loss), rather than the power loss at any specific time point.

For each data-rich network, the time-series of phase residual currents are transformed into a probability density function of the phase residual currents through kernel density estimation (KDE) [88], as given by 4-(2).

$$f(I_n) = \frac{1}{n \cdot h}\sum_{t=1}^{n} K\left(\frac{I_n - I_n(t)}{h}\right) \quad \text{4-(2)}$$

where I_n denotes the phase residual current; $I_n(t)$ is the phase residual current at time t; n denotes the sample size; h denotes the kernel bandwidth. In this paper, the kernel function K is chosen to be the Gaussian kernel [89], given by :

$$K\left(\frac{I_n - I_n(t)}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{I_n - I_n(t)}{h}\right)^2} \quad \text{4-(3)}$$

where *h* is the bandwidth, given by [88]:

$$h = 1.06 \cdot \sigma \cdot n^{-\frac{1}{5}} \quad \text{4-(4)}$$

where σ denotes the standard deviation of the sample data; n denotes the sample size.

For each data-rich network, the probability density function of the phase residual currents is transformed into a CDF. Therefore, there are a total of 800 phase residual current CDFs for the 800 data-rich LV networks.

4.3.2. Clustering

Agglomerative hierarchical clustering and k-means clustering are applied to cluster these 800 phase residual current CDFs into *k* clusters. The reason why I use agglomerative hierarchical clustering and k-means clustering is that they are commonly used classic clustering methods [90], [91]. The agglomerative hierarchical clustering method starts by taking each CDF as its own cluster; then, it generates higher-level clusters by merging clusters with the least dissimilarity between each other until eventually achieving only one cluster [91]. This subsection presents three detailed aspects: 1) distance metrics; 2) the selection of the number of clusters, and 3) the evaluation of clustering results.

Both Euclidean distance (ED) [86] and Jensen-Shannon distance (JSD) [92] are applied to calculate the dissimilarity between any two CDFs.

1) Determine the number of clusters

In this paper, the number of clusters *k* is determined by a bi-objective optimization model. The optimization model aims to minimize the weighted sum of: 1) an overlap ratio; and 2) the relative within-cluster sum of squared distances. The optimization model is given by

$$\begin{aligned} &\min_{k} C \cdot r(k) + s(k) \\ \text{subject to } \quad &2 \le k \le k_{up} = \text{argmax}\, r(k) \\ &k \text{ is an integer} \\ &0 \le r(k) < 1 \\ &0 \le s(k) < 1 \end{aligned} \qquad \text{4-(5)}$$

where C is a weighting factor ($C > 0$); $r(k)$ is the overlap ratio defined in 4-(6); $s(k)$, defined in 4-(7), is the relative within-cluster sum of squared distances as a function of k.

Now this paper defines the overlap ratio $r(k)$. Because this paper estimates the annual imbalance-induced energy loss, which is proportional to the sum of data-rich network's squared phase residual currents over a year, the clustering results are considered "good" if different clusters are distinguishable from each other in terms of their distributions of the sums of squared phase residual currents over a year. In other words, each cluster shall have a distinct distribution of the sum of squared phase residual currents as compared to other clusters. To quantify such a distinctiveness, the overlap ratio is defined in 4-(6).

$$r(k) = n_o(k)/N \qquad \text{4-(6)}$$

where *k* denotes the number of clusters; $r(k)$ is the overlap ratio as a function of *k*; n_o is the number of data-rich networks that have the same sum of squared phase residual currents across different clusters (the shadow area as illustrated in Fig. 4-2). N denotes the total number of data-rich networks. An illustration of the overlap ratio is given in Fig. 4-2.

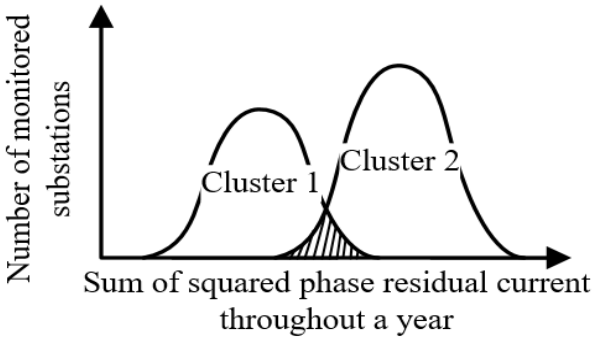


Fig. 4-2 The objective overlap area

The shadow area in Fig. 4-2, i.e., the overlap of the two clusters 1 and 2, represents n_o in 4-(6) – this can be easily extended to *k* clusters. The overlap ratio $r(k)$ is the shadow area divided by the total area of all clusters. When k increases from 2 to the maximum number of clusters (800 in this case), $r(k)$ first increases then decreases to zero. Denote k_{up} as the k value when $r(k)$ reaches the maximum.

Now this paper defines the relative within-cluster sum of squared distances $s(k)$, as given by

$$s(k) = \frac{\sum_{j=1}^{k} \sum_{i\ (x_i \in \text{cluster } j)} (x_i - x_{p,j})^2}{\sum_i (x_i - \bar{x_l})^2} \quad \text{4-(7)}$$

where x_i denotes the ith element in cluster j; $x_{p,j}$ is the prototype of cluster j; $x_{p,j}$ is the medoid of all elements.

2) Evaluate clustering results

After determining the number of clusters *k*, the agglomerative hierarchical clustering process is straightforward. The results show that the agglomerative hierarchical clustering with Euclidean distance yields the least overlap ratio, as compared to k-means with Euclidean distance, *k*-means with Jensen-Shannon distance, and

agglomerative hierarchical clustering with Jensen-Shannon distance. The numerical results and detailed discussions are presented in Chapter 4.5 (case studies). Therefore, the agglomerative hierarchical clustering with Euclidean distance is chosen as the method for clustering the 800 phase residual current CDFs. The clustering output is a cluster label for each data-rich network, indicating which cluster this network belongs to. The medoid of each cluster is selected to be the prototype of this cluster [80].

4.3.3. Classification

Given the clustering outputs, the classification process consists of the following steps: 1) feature vectors (input data for classification) are determined for both the data-scarce and data-rich networks; 2) the feature vectors and cluster labels for the 800 data-rich networks are used to train the classification model by applying multiclass support vector machine (MSVM) and kernel-based Adaptive Boost (kAdaBoost); MSVM and kAdaBoost then classify the data-scarce network to an existing cluster of data-rich networks. The classification results are validated by 10-fold cross-validation.

1) Determine feature vector

Data-scarce networks do not have time-series data, and they account for the majority of the UK's LV networks. They only have data collected once a year. According to [87], this paper suggests that the yearly average currents for three phases ($I_{ava}, I_{avb}\ and\ I_{avc}$) be chosen as the known data for data-scarce networks: 1) DNOs can obtain them directly from existing devices in a low-cost fashion for millions of networks, and these data do not require the deployment of any high-resolution monitoring device; 2) the features derived from these data allow for relatively high classification accuracy.

Given the yearly average phase currents, this paper proposes a feature vector consisting of two features: the virtual average phase residual current value I_{vprc} and virtual average balanced current value I_{vbc}. They can be readily calculated from the yearly average phase currents:

$$I_{vprc} = ({I_{ava}}^2 + {I_{avb}}^2 + {I_{avc}}^2 - I_{ava}I_{avb} - I_{ava}I_{avc} - I_{ava}I_{avc})^{1/2} \quad 4\text{-}(8)$$

$$I_{vbc} = (I_{ava} + I_{avb} + I_{avc})/3 \quad 4\text{-}(9)$$

where $I_{ava}, I_{avb}\ and\ I_{avc}$ denote the yearly average phase currents. Therefore, the feature vector $\mathbf{x}_i = [\ I_{vprc}, I_{vbc}]$ is available for the data-scarce network.

For data-rich networks, the above feature vector can be readily derived from the time-series phase residual current data throughout a year. Therefore, each data-rich network has a cluster ID (this is an output from the clustering stage) as its label and a feature vector $\mathbf{x}_i$. Then, the feature vectors and cluster ID for all data-rich networks and the feature vector for the data-scarce network are used as the input data for the classification stage.

2) Classification

The classification is performed by applying two methods, kAdaBoost and MSVM. The reason for choosing MSVM (which uses the support vector machine as the base classifier) is because, by finding the largest margin to separate different classes, the performance of the support vector machine is widely recognized [93], [94]. kAdaBoost is chosen as a candidate because: 1) it reduces the bias of weak learners by combining the weak learners into a strong learner, and it is shown to be resistant against overfitting [95]; and 2) the Gaussian kernel transformation further improve the classification accuracy.

The kAdaBoost method is a combination of the kernel transformation and the well-

established Adaptive Boost method [95]. It consists of the following steps:

Firstly, a Gaussian kernel transformation is applied to transform the original feature vectors $\mathbf{x}_i$ for all networks i (both data-rich and data-scarce) into a high-dimensional feature space. Such a transformation improves the classification accuracy by up to 2%. The Gaussian kernel is given by [62]:

$$K(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{i,k}) = \exp\left(-\frac{\left\|\boldsymbol{x}_{i,j} - \boldsymbol{x}_{i,k}\right\|^2}{2\sigma^2}\right) \quad \text{4-(10)}$$

where $x_{i,j}$ and $x_{i,k}$ denote the j_{th} and k_{th} elements of network i's feature vector $\mathbf{x}_i$, respectively; σ^2 is the variance.

Secondly, the Adaboost.M2 model takes the transformed feature $K(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{i,k})$ as the input. For Adaboost.M2, it is essentially a "boosting" method that combines a number of weak classification models ("weak models") into a strong classification model ("strong model") [95]. The strong model is given by [92]:

$$H(x) = \operatorname{argmax} \sum_{t=1}^{T} h_t(x, y) \log \frac{1}{a_t} \quad \text{4-(11)}$$

where h_t is the weak model; a_t denotes the weight parameter. The well-established algorithm of AdaBoost.M2 is detailed in [95].

The MSVM is the multiclass support vector machine [62], [26]. The MSVM is essentially a one-versus-one framework that extends the support vector machine (a binary classifier) into a multiclass classifier [81]. For each binary classification sub-problem, the support vector machine aims to find a separating hyperplane in the high-dimensional feature space (as a result of the Gaussian kernel transformation of the feature vectors) to separate the two classes with the maximum margin [94]. The support vector machine essentially solves an optimization problem, as given by [96].

$$\min_{\boldsymbol{\omega},\, b} \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{N_t} e_i$$
$$\text{subject to } y_i(\boldsymbol{\omega}^T \cdot \varphi(\mathbf{x}_i) + b) \geq 1 - e_i \qquad \text{4-(12)}$$
$$e_i \geq 0$$
$$y_i \in \{-1, 1\}$$

where $\boldsymbol{\omega}$ and b are the coefficient vector and the interception term, respectively; y_i is the label for training example i; $\varphi(\mathbf{x}_i)$ is the transformed feature vector in the high-dimensional space for training example i. $C\sum_{i=1}^{N_t} e_i$ is the regularization term that reduces the generalization error, where C denotes the penalty coefficient; N_t denotes the total number of training examples; e_i represents the infringement an outlier causes. The algorithm of MSVM is detailed in [93].

The classification process is validated by 10-fold cross-validation. This is a well-established, popular validation method. It is detailed in [97], [98].

The classification results from the two methods are compared with each other in the case studies. Given the clustering and classification model trained and the data-scarce network, the output of the classification stage is the cluster to which this network is classified.

4.4. Imbalance-induced energy loss range estimation

The classification stage in Chapter 4.3.3 classifies the data-scarce network into an established cluster derived in Chapter 4.3.2. The maximum range of the imbalance-induced energy loss for this cluster is then derived. This range is then narrowed down to a confidence range by applying Chebyshev's inequality formula. This confidence range is where the imbalance-induced energy loss of the data-scarce network falls at a

predefined confidence level, as cross-validated in Chapter 4.5. Detailed steps are given below.

Firstly, the imbalance-induced energy losses for these data-rich networks are calculated for two different earthing systems, TN-C and TN-S. The TN-C earthing system is presented in Fig. 4-3 [38] :

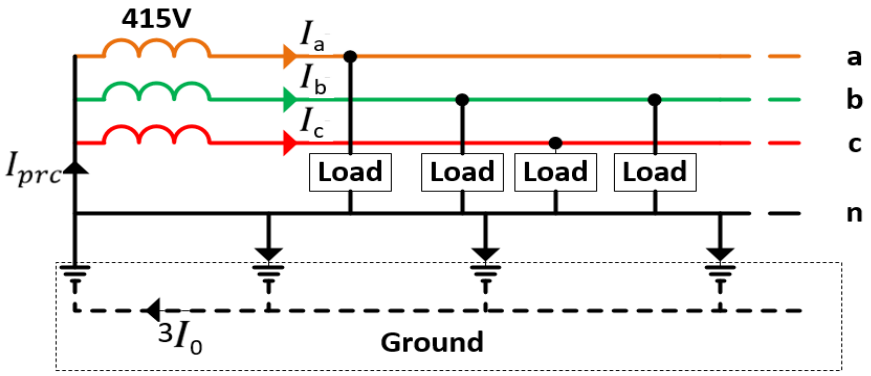


Fig. 4-3 The TN-C earthing system

For the TN-C earthing system, I_{prc} is the phase residual current that flows into the transformer neutral point from the ground [38]. The imbalance-induced power loss is given by

$$P_{loss} = I_{prc}(t)^2 R_g \quad \text{4-(13)}$$

where I_{prc} denotes the phase residual current; R_g is the equivalent ground resistance, which is 0.0953 (Ω/km) · Length (km).

The TN-S earthing system is shown in Fig. 4-4 [38]:

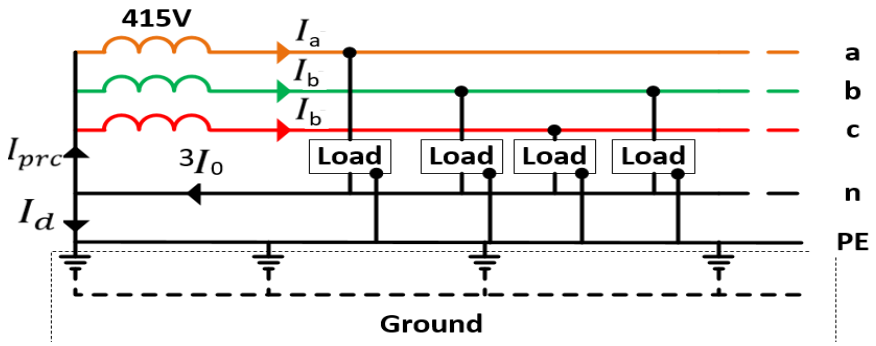


Fig. 4-4 The TN-S earthing system

For the TN-S earthing system, the protective wire and the neutral wire are separate conductors. When there is phase imbalance, the phase residual current I_{prc}, flows into the transformer neutral point through the neutral conductor. Therefore, the imbalance-induced power loss is given by

$$P_{loss} = I_{prc}(t)^2 R_n \quad \text{4-(14)}$$

where I_{prc} denotes the phase residual current; R_n denotes the neutral wire resistance.

Secondly, given that the clustering stage in Chapter 4.3.2 has already clustered the 800 data-rich networks into *N* clusters, the maximum range $[E_{lossmin}, E_{lossmax}]$ of the imbalance-induced energy loss for each cluster is derived, where $E_{lossmin}$ and $E_{lossmax}$ denote the minimum imbalance-induced energy loss and the maximum imbalance-induced energy loss, respectively.

The above maximum range is sensitive to outliers. To counter the impact of outliers, the maximum range of the imbalance-induced energy loss for each cluster is narrowed down to a confidence range by applying Chebyshev's inequality formula [82]. In industry, a common practice is to remove 1 – 2% of the observed data close to the range boundaries [99], assuming that the data follow a Gaussian distribution. The reason why I choose Chebyshev's inequality formula is that, unlike other methods, it

does not require that the data follow any particular classic distribution (e.g. Gaussian distribution). In this paper, the imbalance-induced energy loss results for any cluster of data-rich networks are not assumed to follow any particular classic distribution. Therefore, Chebyshev's inequality formula is suitable in this case. Chebyshev's inequality formula states that the probability of a random variable falling beyond $k\sigma$ from its mean is less than $1/k^2$, as given by

$$\text{Prob}\left(|x-\mu| \geq k\sigma\right) \leq 1/k^2 \qquad \text{4-(15)}$$

where x is the random value of the imbalance-induced energy loss; μ denotes the expectation of the imbalance-induced energy loss; σ is the standard deviation of the imbalance-induced energy loss; k is the coefficient. Reference [100] suggests that the coefficient k be set as 3 to remove outliers, which means that the values falling in the interval $[\mu-3\sigma, \mu+3\sigma]$ has a confidence level of 89%.

The confidence range corresponds to removing 11% of data from the original cluster by Chebyshev's inequality method. An illustration of the "tail cutting" effect is shown in Fig. 4-5.

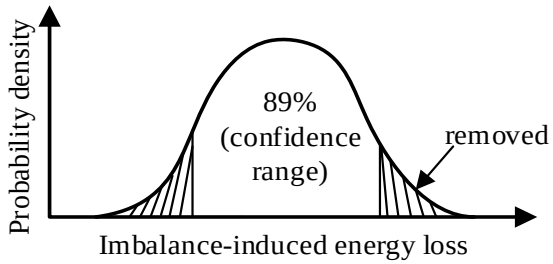


Fig. 4-5 The distribution of example imbalance-induced energy loss for cluster *i*

To implement, the distance between the imbalance-induced energy loss of each data-rich network and the average imbalance-induced energy loss of each cluster *i* is calculated. Then, 11% of the data-rich networks in cluster *i* with larger distances than

the rest are removed. The resulting range of the imbalance-induced energy loss is the 89% confidence range of imbalance-induced energy loss for cluster *i*.

The choice of the 89% confidence level for the range estimation is validated by applying 10-fold cross-validation. For each cluster of *n* data-rich networks, *n* number of imbalance-induced energy loss values are randomly divided into 10 groups of equal size. One of the ten groups of data-rich networks is retained as the validation group; the other 9 groups form a large training group to build a distribution of the imbalance-induced energy loss values. This distribution is narrowed down to the 89% confidence range by applying Chebyshev's inequality formula. Then, the percentage of the validation samples (the imbalance-induced energy loss values within the validation group) that fall within the distribution is calculated. This process repeats until every group has served as the validation group once. This process outputs 10 values, i.e. the percentages of the validation samples falling within the distribution. These 10 values are averaged, and it is found that the average value is close to 89%. In this way, the choice of the 89% confidence level is validated.

The resulting estimation error of the imbalance-induced energy loss is given by

$$error = |AL - EML|/AL \qquad 4\text{-}(16)$$

where *AL* denotes the actual imbalance-induced energy loss (IIBL) of the LV networks; *EML* is the mean value of the estimated range of the imbalance-induced energy loss.

4.5. Case studies

This section presents the numerical results. The clustering and classification results are given in Chapters 4.5.1 and 4.5.2, respectively. The imbalance-induced energy losses are calculated in Chapter 4.5.3. A discussion is presented in Chapter 4.5.4.

4.5.1.Clustering

The first step of clustering is to determine the number of clusters by solving the bi-objective optimization problem in 4-(5). TABLE 4-1 presents the overlap ratio $r(k)$ for different numbers of clusters k.

TABLE 4-1

OBJECTIVE OVERLAP RATIO COMPARISON

Number of clusters	$r(k)$ under the ED metric	$r(k)$ under the JSD metric
6	3.2%	9.8%
7	3.2%	9.8%
8	3.45%	10.1%

In TABLE 4-1, $r(8) > r(7) = r(6)$. $k = 7$ is preferred over $k = 6$ because the former corresponds to a lower sum of within-cluster errors. Therefore, the number of clusters k is chosen to be 7 for both JSD and ED metrics.

Given the number of clusters $k = 7$, the second step is to perform the clustering process using both k-means and hierarchical clustering methods, based on JSD and ED distance metrics. The results are presented in TABLE 4-2 for comparison.

TABLE 4-2

CLUSTERING METHOD COMPARISON

		$\mathrm{r}(k)$	Hierarchical Cophenet
Hierarchical clustering	JSD	9.8%	0.7733
	ED	3.2%	0.7845
K-means clustering	JSD	22.%	
	ED	10.3%	

In TABLE 4-2, the Hierarchical cophenet denotes the cophenet correlation coefficient for the Hierarchical cluster tree, indicating how faithfully the tree represents the dissimilarities among observations (the larger, the better). Hierarchical clustering with

the ED distance metric yields the lowest overlap ratio and a higher cophenet – this combination is therefore chosen for clustering.

Fig. 4-6 and Fig. 4-7 visualize how distinguishable the seven clusters are under: 1) hierarchical clustering with ED metric; 2) hierarchical clustering with JSD metric; 3) k-means with ED metric; and 4) k-means with JSD metric.

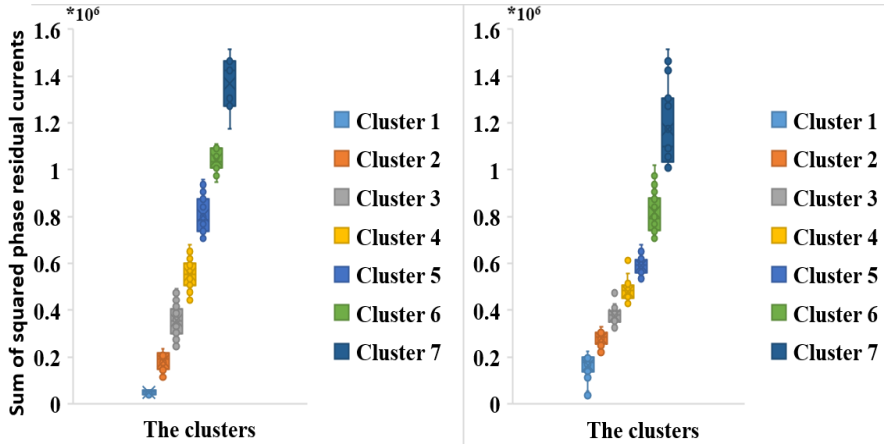


Fig. 4-6 Hierarchical (left) and K-means (right) clustering results with ED metric

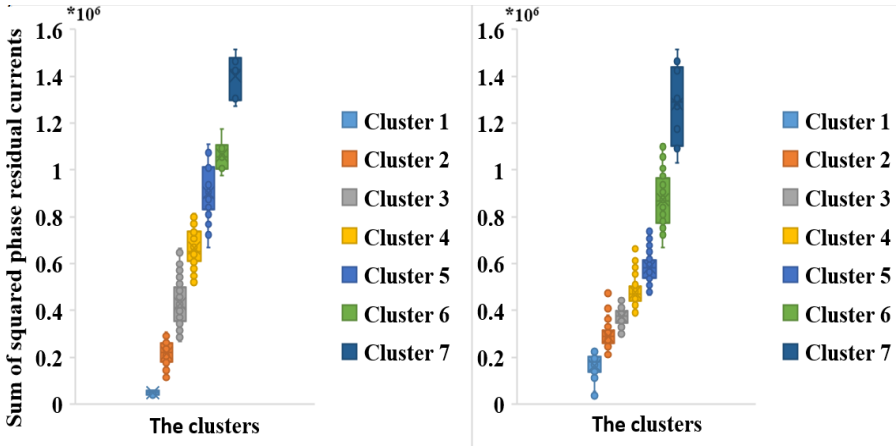


Fig. 4-7 Hierarchical (left) and K-means (right) results with JSD metric

In these diagrams, each cluster is resembled as a bar. Fig. 4-6 and Fig. 4-7 show that hierarchical clustering with the ED distance metric yields the most distinguishable seven clusters as compared to other methods.

The phase residual current CDFs of the data-rich networks within each cluster are plotted as a heat map in Fig. 4-8.

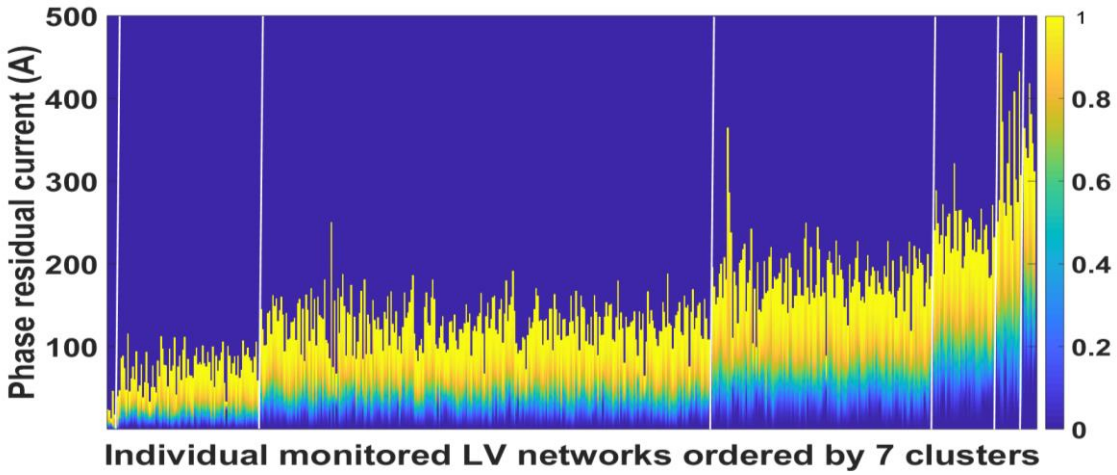


Fig. 4-8 The heat map of the squared phase residual current CDFs of the data-rich networks within each cluster

In Fig. 4-8, the diagram is separated into seven intervals by six vertical white lines, where each interval corresponds to a cluster (from Cluster 1 in the left to Cluster 7 in the right). Each blue-yellow vertical line represents the phase residual current CDF of a data-rich network belonging to the cluster. Each red vertical line represents each cluster's prototype. This figure demonstrates that each cluster has its own phase residual current CDF tendency, which is distinctive from other clusters. In addition, Cluster 1 accounts for 1.09% of the data-rich networks in this study; Clusters 2 – 7 account for 15.25%, 49%, 23.96%, 6.72%, 2.72%, and 1.27% of the data-rich networks, respectively.

4.5.2. Classification

According to Chapter 4.3.2, the virtual average balanced current and virtual average phase residual current are the features used for classification in this sub-section. This

feature is derived from yearly average currents of three phases (I_{ava}, I_{avb} and I_{avc}), recorded once a year by a relay protection metering function. The distribution of the features for each cluster is plotted in Fig. 4-9.

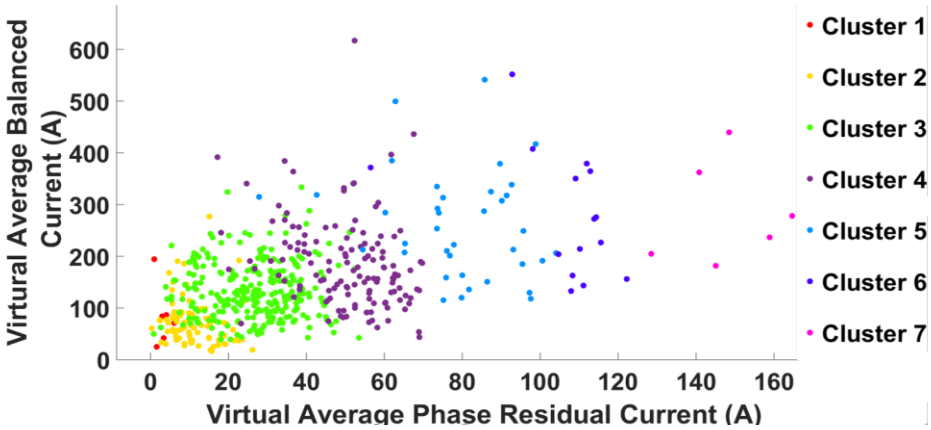


Fig. 4-9 Data-rich networks' feature distribution

Fig. 4-9 shows that the features for different clusters overlap to a large extent. This overlap reflects the data scarcity, i.e., the available feature is rather limited.

From case studies, I find that the Gaussian-kernel-based MSVM and kAdaBoost achieve higher classification accuracies than alternative classification methods such as k-Nearest Neighbmys (KNN) and decision tree. The comparison of the classification accuracies is presented in Fig. 4-10.

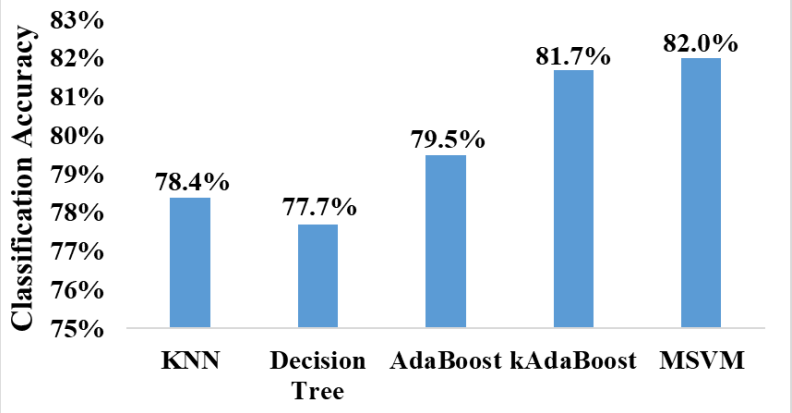
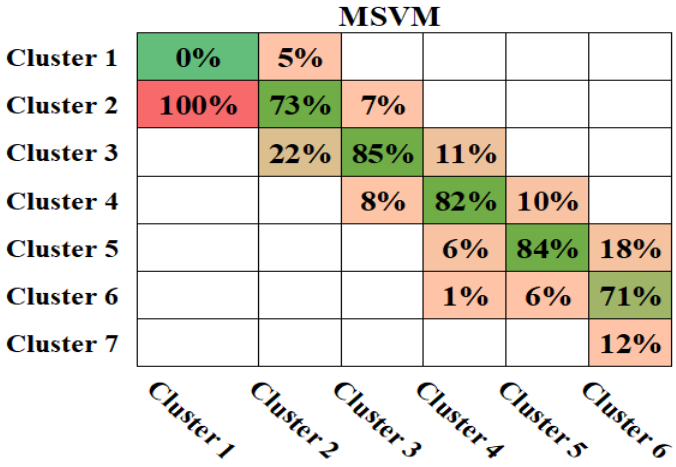


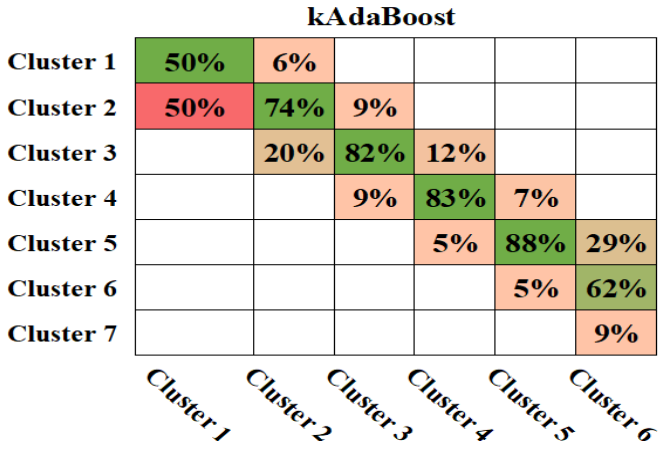
Fig. 4-10 The classification results comparison of different methods

From Fig. 4-10, the MSVM achieves the highest classification accuracy of 82%, followed by kAdaBoost which achieves a classification accuracy of 81.7% and adaptive boost (AdaBoost) which achieves 79.5% accuracy. KNN and decision tree achieve 78.4% and 77.7% accuracies, respectively. In comparison, a blind guess would give an accuracy of only 14.29%.

The confusion matrices for the classification results by MSVM and kAdaBoost are presented in Fig. 4-11.



(a)



(b)

Fig. 4-11 Confusion matrices for the MSVM and kAdaBoost methods

The confusion matrices in Fig. 4-11 demonstrate the classification accuracies in details. For instance, for the MSVM classification, column two shows that the data-scarce network which should be classified into Cluster 2 has 5% probability of being misclassified into Cluster 1, 22% probability of being misclassified into Cluster 3.

Both classification methods only require virtual average balanced current and virtual average phase residual current, derived from the yearly average currents of three phases (I_{ava}, I_{avb} and I_{avc}), as the feature from data-scarce LV networks. This means it can be implemented in a cost-effective manner using existing devices only.

For example, a data-scarce network has the yearly average phase currents $[I_{ava}, I_{avb}, I_{avc}] = [219.1\text{A}, 182.4\text{A}, 224.1\text{A}]$. These data are transformed into a feature vector $\mathbf{x}_i = [I_{vbc}, I_{vprc}] = [208.5\text{A}, 39.4\text{A}]$. Given this feature vector, this data-scarce network is classified into Cluster 4 by applying either MSVM or kAdaBoost.

4.5.3.Imbalance-induced energy losses estimation

The resistance of the path on which the phase residual current flows is affected by many factors, including the length of the path, the resistivity of the cables and the ground, ambient condition, and the topology, etc. To account for the complicated nature, this paper considers multiple scenarios on the resistance and estimates the imbalance-induced energy losses for these scenarios. According to [101], the length of the UK's LV networks normally ranges from 0.9 km to 2.1 km; the resistivity of the ground is 0.0953 Ω/km; the resistivity of the neutral conductor ranges from 0.168 Ω /km to 0.320 Ω /km. Therefore, for TN-C earthing system, the ground resistance R_g varies from 0.0858 Ω to 0.2001 Ω; for TN-S earthing system, the neutral conductor resistance R_n varies from 0.1512 Ω to 0.6720 Ω;

For the TN-C earthing system, the confidence range of the imbalance-induced energy losses for each cluster is plotted in Fig. 4-12:

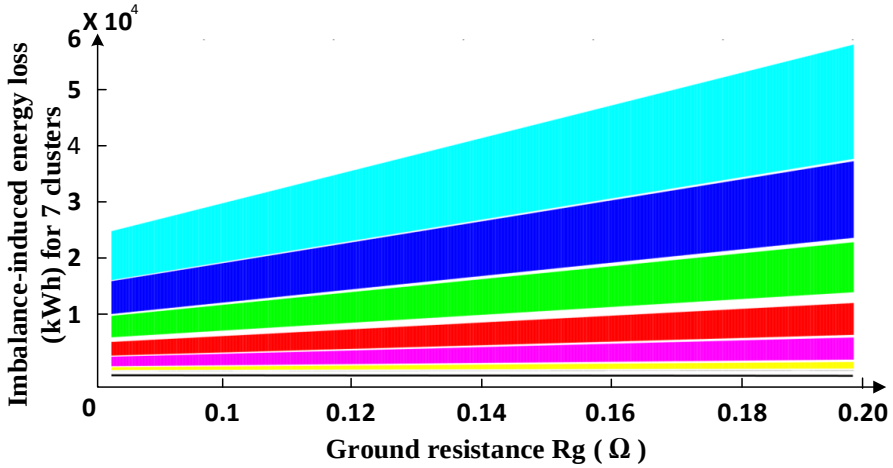


Fig. 4-12 The confidence range of the imbalance-induced energy losses of TN-C earthing system for the clusters

For example, when the ground resistance is 0.143 Ω (a length of 1.5 km, which is the average length of the UK's LV networks), for Cluster 1, the confidence range of the imbalance-induced energy losses is [54 kWh, 76 kWh] per year. The confidence ranges of the imbalance-induced energy losses for Clusters 2 – 7 are [328 kWh, 1,163 kWh], [1,457 kWh, 4,271 kWh], [4,601 kWh, 8,638 kWh], [10,005 kWh, 16,345 kWh], [16,904 kWh, 26,615 kWh], and [26,914 kWh, 41,405 kWh] per year, respectively.

Given an estimation of 900,000 networks throughout the UK and an average electricity price of £ 0.18/kWh, the phase imbalance situation causes 3.01×10^6 to 6.02×10^6 MWh of imbalance-induced energy losses each year, worth £451.2m to £903.0m per annum.

For TN-S earthing system, the neutral conductor resistance R_n varies from 0.1512 Ω to 0.6720 Ω. The confidence range of the imbalance-induced energy losses for each cluster is plotted in Fig. 4-13:

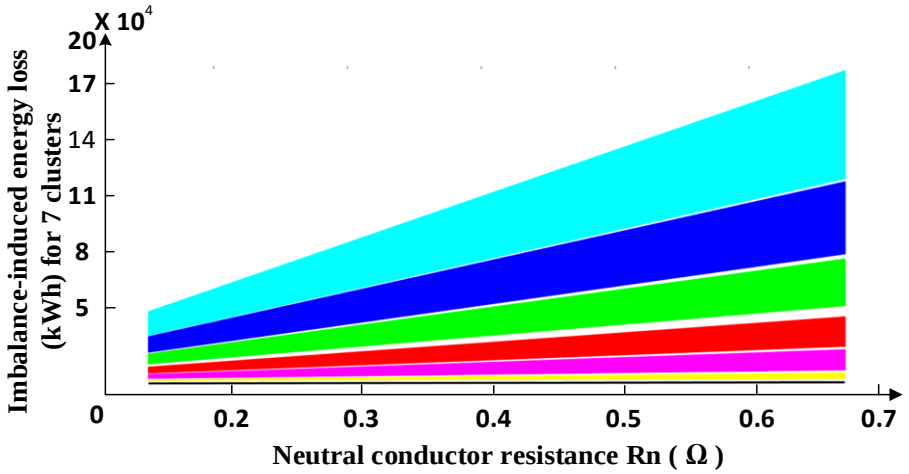


Fig. 4-13 The confidence range of the imbalance-induced energy losses of TN-S earthing system for the clusters

For example, if the neutral conductor resistance is 0.252 Ω (with a length of 1.5 km and a resistivity of 0.163 Ω/km), for Cluster 1, the confidence range of the imbalance-induced energy losses is [94 kWh, 135 kWh] per year. The confidence ranges of the imbalance-induced energy losses for Clusters 2 – 7 are [578 kWh, 2,050 kWh], [2,569 kWh, 7,531 kWh], [8,114 kWh, 15,233 kWh], [17,644 kWh, 28,824 kWh], [29,809 kWh, 46,934 kWh], and [47,461 kWh, 73,016 kWh] per year, respectively.

Given an estimation of 900,000 networks throughout the UK and an average electricity price of £ 0.18/kWh, the phase imbalance situation causes 5.3×10^6 to 1.06×10^7 MWh of imbalance-induced energy losses each year, worth £795.3m to £1,592m per annum.

This paper applies a 10-fold cross-validation to validate the confidence range of the annual imbalance-induced energy loss. The cross-validation results show that 9% of the data-rich networks that belong to Cluster 2 fall beyond the confidence range of Cluster 2; 11%, 12%, and 11% of the data-rich networks that belong to Clusters 3, 4, and 5 falls beyond the respective confidence range of the cluster. Clusters 1, 6, and 7 have 5, 15, and 6 data-rich networks, respectively – too few networks that it is not

suitable to remove any data from them. Therefore, the confidence ranges of Clusters 1, 6, and 7 are the maximum range of these clusters.

The same example in Chapter 4.5.2 is used. Its estimated imbalance-induced energy loss is within a confidence range of [1,074 kWh, 2,131 kWh] per year, with a confidence level of 89%.

TABLE 4-3 presents examples showing the estimation errors:

TABLE 4-3

EXAMPLE OF THE CCRE ESTIMATION ERROR

	I_{vprc} (A)	I_{vbc} (A)	Correct cluster	Actual IIBL (kWh)	Classified cluster	Estimated range of IIBL (kWh)	Estimation error
1	87.3	324	5	24,520	6	29,809 – 46,934	61.15%
2	19.7	336	3	3,096	2	577 – 2,050	55.91%
3	98.0	407	6	38,350	5	17,644 – 28,824	40.92%
4	17.8	38.1	2	1,692	2	577 – 2,050	19.33%
5	59.9	177	4	9,580	4	8,114 – 15,233	18.29%
6	145	181	7	54,386	7	47,461 – 73,016	13.79%

In TABLE 4-3, the first three examples are classified into the wrong clusters, resulting in substantial errors of more than 40%. The last three examples are classified to the correct clusters, resulting in errors of less than 20%.

4.5.4.Discussion

To estimate the imbalance-induced energy loss, the proposed CCRE approach only requires the yearly average phase currents as the feature from data-scarce networks. This feature can be readily obtained from existing LV networks. This renders the CCRE approach applicability to the majority of the UK's LV networks that are data-scarce, without the need for high-resolution monitoring devices on neutral wires.

In this paper, the 800 CDFs of the phase residual current I_{prc} are used as the input data for clustering. The energy loss is proportional to the square of the phase residual current, i.e. I_{prc}^2. However, the reason why the CDFs of I_{prc} are used as the input data instead of the CDFs of I_{prc}^2 is because the latter would increase the data dispersion from 0 – 300 to 0 – 90,000. This expands the range of the CDFs to a level too wide for clustering. Furthermore, the clustering results show that the former results in an overlap ratio as low as 3.2%, whereas the latter results in an overlap ratio of more than 20%. Therefore, the former is much better than the latter as the input data for clustering.

The CCRE approach is designed to be generic. To apply the CCRE approach to other countries, it would require the following two groups of input data for the country in question: 1) the time-series phase current data monitored throughout a year from at least hundreds of data-rich LV networks (these data are used as the training data); and 2) the yearly average phase currents for the data-scarce network (these limited data are called the feature). The more representative the training data are, the more accurate the estimated phase residual current for the data-scarce network is.

This paper considers phase residual current profiles and there is a fundamental difference between a load profile and a phase residual current profile. The former depends on the number of customers and types of customers, whereas the latter depends on how evenly (or unevenly) customers are allocated across the three phases. Because urban, suburban, and rural areas have very different customer densities and types of customers, their load profiles are different – the classification of load profiles into these four areas is justified. However, different types of areas may have the same degree of phase imbalance, i.e. customers in these areas are allocated in the same uneven fashion, thus resulting in similar phase residual current profiles. On the other hand, two networks in the same type of areas (e.g. urban) may have very different degrees of phase imbalance, resulting in vastly different phase residual current profiles.

Therefore, the division into urban, suburban, and rural areas is not applicable in this paper.

There can be full current measurements from high-voltage (132 kV / 33 kV) and medium-voltage (33 kV / 11 kV) distribution substations as well as customer billing data. However, these measurements are not normally available from low-voltage (11 / 0.415 kV, LV) substations downwards (inclusive), because of the prohibitively high cost to monitor millions of LV networks. Furthermore, even if smart meter data were available for all customers (which is not the case in the UK now), which phase each customer is connected to is still unknown [102], [103]. Because of the above field limitations, state estimation cannot be performed for LV networks.

The load loss factor method is popular for calculating energy losses. However, it is not suitable in this paper, because it requires the average phase residual current and the maximum phase residual current as the input data, which are not available for data-scarce LV networks. Furthermore, the load loss factor is suggested to be updated every month to minimize the error of the estimation [44]. For the data-scarce networks, the cost to update the load loss factors for 900,000 LV networks every month would be unimaginably high.

Increasing available features would improve the accuracy of the classification. If the sum or average of the phase residual currents over a year were known for data-scarce networks, the CCRE approach would achieve an accuracy of 96.8%, much higher than if only the average phase residual currents are known. However, increasing features pose more requirements on the monitoring of the LV networks, resulting in more costs.

Phase imbalance causes two costs: 1) the imbalance-induced energy loss; and 2) the additional network investment cost. These two costs are required to be estimated for a

cost benefit analysis of any phase balancing project. This paper finds out whether the 1st cost element is significant or not and how significant it is for both highly phase-imbalanced LV networks and not-so-imbalanced LV networks. Furthermore, this paper calculates the 1st cost for one year only. In reality, this cost occurs year by year until the three phases are fully balanced. Future work will be to perform a full cost-benefit analysis for phase balancing solutions considering the above two benefits together, the lack of data in LV networks, and the uncertainty associated with the phase balancing capability.

4.6. Discussions on increasing visible data for improving the estimation accuracy

The section discusses increasing visible data from data-scarce LV networks improves the estimation accuracy of the CCRE methods. This section is the expanding contents of this discussion.

This chapter utilises minimal data, the yearly average phase current, from data-scarce LV networks for estimating their imbalance-induced residual energy losses with over 80% accuracy. These data are commonly existing data for DNOs with minimal accessing costs. Having more visible data could increase the estimation accuracy. For example, supposing DNOs collect the yearly phase residual current by deploying an additional current meter on the residual path of LV networks. The feature vector for data-scarce LV networks is given by:

$$\mathbf{x}_i = [\, I_{vprc}, I_{vbc}, \overline{I_{rc}}] \qquad 4\text{-}(17)$$

where I_{vprc}, I_{vbc} are defined in 4-(8) and 4-(9) respectively; I_{vrc} gives the average phase residual current.

Give the new feature vector, and the same example as that in the case studies, the average accuracy for estimating the imbalance-induced residual energy losses ascends to 96.8%. However, it is costly to deploy such additional current meters for collecting the average phase residual current for a total number of over 900,000 LV networks in the UK. Therefore, for the industry, DNOs should make a trade-off between estimation accuracy and costs. The method in this chapter delivers over 80% accuracy with minimal data accessing costs.

4.7. Conclusions

This paper addresses an unsolved problem faced by utility companies, i.e., estimating imbalance-induced energy losses for data-scarce low voltage (415V, LV) networks with only the yearly average phase currents data.

The 800 LV data-rich networks with full time-series of phase currents data are clustered into 7 clusters, where each cluster represents networks of similar phase residual current profiles. Then, at the classification stage, cross-validation results show that nearly 82% of the data-scarce networks are classified to the correct clusters. The confidence interval of the imbalance-induced energy loss for the data-scarce network is derived at a confidence level of 89%. The proposed methodology enables distribution network operators to evaluate a key cost of phase imbalance. This cost serves as a necessary input for the appraisal of the benefit from phase balancing.

4.8. Appendix

The phase residual current (calculated by equation 4-(1)) is the vector sum of the phase currents:

$$\overrightarrow{I_{prc}} = \overrightarrow{I_a} + \overrightarrow{I_b} + \overrightarrow{I_c} \quad \text{4-(18)}$$

In the absence of phasor measurements, it is assumed that the phase currents are 120° apart from each other. Therefore,

$$\begin{aligned}\overrightarrow{I_{prc}} = {} & I_a \cos 0^\circ + jI_a \sin 0^\circ + I_b \cos -120^\circ + jI_b \sin -120^\circ + I_c \cos 120^\circ \\ & + jI_c \sin 120^\circ \\ = {} & (I_a - \frac{1}{2}I_b - \frac{1}{2}I_c) + j(\frac{\sqrt{3}}{2}I_c - \frac{\sqrt{3}}{2}I_b)\end{aligned} \quad \text{4-(19)}$$

$$\begin{aligned}|I_{prc}| &= \sqrt{\left(I_a - \frac{1}{2}I_b - \frac{1}{2}I_c\right)^2 + \left(\frac{\sqrt{3}}{2}I_c - \frac{\sqrt{3}}{2}I_b\right)^2} \\ &= \sqrt{I_a^2 + I_b^2 + I_c^2 - I_aI_b - I_bI_c - I_aI_c}\end{aligned} \quad \text{4-(20)}$$

where I_{prc} is the phase residual current; I_a, I_b and I_c denote the magnitudes of the phase currents.

4.9. Chapter summary

This chapter developed one novel approach to address one unsolved engineering question for data-scarce LV networks: assessing imbalance-induced residual energy losses. This approach delivers 82% estimation accuracy with range estimation. Considering this energy loss acts as a significant part in improving network planning methods or making long-term phase balancing investment decisions, delivering a range estimation is a more robust answer than point estimation.

This approach is validated using data from 800 data-rich networks within the WPD's business area. For other DNOs or other countries, before implementing these two approaches, this thesis recommends that: 1) collecting time-series data from a small portion, but representative, data-rich networks (no less than 800) within their own business area; 2) following my approaches' flowcharts to train their own models for

estimating imbalance-induced phase energy loss and imbalance-induced residual loss. It should be noted that the machine learning methods used in my approaches are not always the best. Developing a statistical approach should always uphold the “no free lunch theorem” [104], i.e., there is no one fits all solution. Other DNOs within or out of the UK should adopt different classical machine learning methods to train their models and compare their estimation accuracy to determine the most appropriate method. For example, the determined classification tool is k-Adaboost in this chapter, where the alternative tools are ordinary linear classification, MSVM, reinforcement learning and so forth. DNOs should always select the tool that delivers the most outstanding estimation accuracy.

The developed statistical approach has limitations as other data-driven approaches should have. First, the results of statistical approaches are very sensitive to the data itself. When implementing my approaches in field works, the input data-rich networks should be representative as I have in case studies – the data-rich networks should have a good mixture of urban, suburban and rural LV networks and a good mixture of domestic, commercial and industrial customers. Second, the trained model should be updated each year to ensure the results are accurate.

Chapter 5.

Guiding phase swapping for data-scarce LV networks

Chapter contents:

This chapter develops one original methodology to making phase swapping guidance for data-scarce LV networks, without the requirement of deploying additional monitoring devices and customer' smart meter access.

5.1. Chapter summary

Phase swapping is a classic method to address phase swapping by moving customers from one phase to another. Substantial references have studied how to making phase swapping strategies, assuming the network topologies, customer's smart meter data, and network's substation-side time-series data are known. However, in reality, these data are unknown for the majority of LV networks. For these data-scarce LV networks, it raises a problem: how to develop phase swapping guidance without deploying additional monitoring devices.

This chapter, for the first time, develops a statistical approach to get around the problem. In detail, first, given a set of data-rich LV networks (with year-round substation-side time-series phase current data), this approach uses non-negative matrix factorization [105],[106] (NMF), a blind source separation approach, to extract a set of typical load profiles. At this stage, phase swapping guidance is developed for these data-rich networks, as well as their rebalancing potentials (potential reduction of phase imbalance degree by implementing the derived phase swapping guidance). Second, a rapid screen model is developed by learning the relationship between the features of the data-rich networks and their calculated rebalancing potentials. It should be noted that the selected feature for these data-rich networks should also exist in data-scarce networks. Then, this rapid screening model estimates rebalancing potentials for data-scarce LV networks and identifies data-scarce LV networks with high rebalancing potentials. Third, for LV networks with high rebalancing potentials, I suggest that DNOs should collect one day's phase current data to ensure developing a credible phase swapping guidance. Case studies reveal that the statistical approach produces effective phase swapping guidance, which reduces the phase imbalance degrees for 99% of the data-scarce LV networks, and the maximum reduction is 0.35. Moreover, the reduction of phase imbalance degree for data-scarce LV networks is only

14.3% lower than that for data-rich networks.

The rest of this chapter is cited from the author's published article in IEEE transactions on Power systems [107]. The chapter is organised in an alternative-based format, where the indices, equations, tables, figures and titles are numbered independently.

Statement of Authorship

<table>
<tr><td colspan="4">This declaration concerns the article entitled:</td></tr>
<tr><td colspan="4">A Statistical Approach to Guide Phase Swapping for Data-Scarce Low Voltage Networks</td></tr>
<tr><td colspan="4">Publication status (tick one)
Draft manuscript [ ] Submitted [ ] In review [ ] Accepted [ ] Published [√]</td></tr>
<tr><td>Publication details (reference)</td><td colspan="3">Fang, L., Ma, K., & Zhang, X. (2020). A Statistical Approach to Guide Phase Swapping for Data-Scarce Low Voltage Networks. IEEE Transactions on Power Systems, 35(1), 751-761. https://doi.org/10.1109/TPWRS.2019.2931981</td></tr>
<tr><td colspan="4">Copyright status (tick the appropriate statement)
I hold the copyright for this material [ ] Copyright is retained by the publisher, but I have been given permission to replicate the material here [√]</td></tr>
<tr><td>Candidate's contribution to the paper (provide details, and also indicate as a percentage)</td><td colspan="3">The candidate contributed to / considerably contributed to / predominantly executed the…
Formulation of ideas:
● 80%
● Defining the real problem in making phase swapping strategies for massive LV networks and the solution's implementation challenges, guided by Dr Kang Ma.
Design of methodology:
● 90%
● Customising a statistical approach combining signal processing, rapid-screening model fitting and the least square model to make phase swapping guidance for data-scarce LV networks, guided by Dr Kang Ma.
Experimental work:
● 100%
Presentation of data in journal format:
● 80%
● Organising and writing this article, revised by Dr Kang Ma, Dr Xinsong Zhang</td></tr>
<tr><td>Statement from Candidate</td><td colspan="3">This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.</td></tr>
<tr><td>Signed</td><td>Lurui Fang</td><td>Date</td><td>30/07/2021</td></tr>
</table>

A Statistical Approach to Guide Phase Swapping for Data-Scarce Low Voltage Networks

Lurui Fang, *Student Member, IEEE,* Kang Ma, *Member, IEEE*, Xinsong Zhang

Abstract—Phase swapping, which rebalances the unbalanced three-phase low voltage (LV, 415V) networks, improves network efficiency by reducing capacity waste and energy losses. A key challenge against phase swapping is that the majority of LV networks are data-scarce, i.e., there is a general lack of data in LV networks. In light of this, this paper proposes a new statistical approach to develop phase swapping guidance for data-scarce LV networks with neither time-series network measurements nor customer metering data. Firstly, given a set of data-rich LV networks (with time-series phase currents data collected at LV substations throughout a year), typical load profiles and their weights in each of the three phases are extracted by applying a non-negative matrix factorization method. Then, phase swapping guidance is developed for data-rich LV networks along with their rebalancing potentials (rebalancing potentials refer to the reduction of phase imbalance degree). Secondly, a rapid screening model is developed to efficiently identify the data-scarce LV networks with high rebalancing potentials. Phase swapping guidance are then developed for these data-scarce networks with high rebalancing potentials. Case studies reveal that the statistical approach produces effective phase swapping guidance, which reduces the phase imbalance degrees for 99% of the LV networks and the maximum reduction is 35%. Validation results show that the average reduction of the phase imbalance degree for data-scarce networks is only 14.3% less than that for data-rich networks.

Index Terms—low voltage, phase imbalance, phase balancing, phase swapping, power distribution, three-phase power, statistical approach

Nomenclature

$DPIB_{ori}$	The original phase imbalance degree
$DPIB_{Bal}$	The phase imbalance degree after rebalancing
$DPIB_v$	The virtual phase imbalance degree
$DPIB_{sba}$	The phase imbalance degree after rebalancing for the validation sample
e_v	The rebalancing error
$\mathbf{H}$	The matrix of weighting factors
$\mathbf{H}_{net}$	The weighting factor matrix for a data-rich network
$\mathbf{H}_{\mathrm{bm}}$	The balanced weighting factor matrix
$\mathbf{H}_{\mathrm{ds}}$	The weighting factor matrix for a data-scarce network
$\mathbf{H}_{\mathrm{bms}}$	The balanced weighting factor matrix for a data-scarce network
$h_{\emptyset}$	The weighting factors in phase $\emptyset$ $(\emptyset\epsilon\{\mathrm{a},\mathrm{b},\mathrm{c}\})$ of the data-rich network
$\mathbf{I_{PI}}$	The input phase current matrix
$\mathbf{I}_{SB}$	the rebalanced time-series phase current data for a validation sample
$I_{p\emptyset,i}(t)$	The time-series phase current data for phase $\emptyset$ $(\emptyset\epsilon\{\mathrm{a},\mathrm{b},\mathrm{c}\})$ of the i_{th} data-rich LV networks
$I_{P\emptyset B}$	The rebalanced time-series phase current profiles
I_{ya}, I_{yb}, I_{yc}	The yearly average phase current data
$\mathrm{I}_{o\emptyset}(t)$	The one day's phase current data for phase $\emptyset$ $(\emptyset\epsilon\{\mathrm{a},\mathrm{b},\mathrm{c}\})$
n_t	The length of the time series phase current data
n_p	The total number of phases for all data-rich LV networks
n_{net}	The number of data-rich LV networks
n_d	The number of the constituent load profiles
n_{dt}	The length of one-day' time-series phase current data
n_{dy}	The number of days throughout a year
$\mathbf{SPSS}$	The statistical phase swapping matrix

5.2. Introduction

Phase imbalance causes significant consequences to low voltage networks (415V, LV), e.g. extra energy losses [7], [8], additional reinforcement cost [77], risks of network nuisance tripping (because of a high zero-sequence current) [9], risks of network overloading [108], and possible damages to induction motors because of voltage imbalance [10], [109]. Phase swapping is a natural way to rebalance the three phases and resolve the above problems [9], [108], [2]. However, developing mass-scale guidance for phase swapping remains a challenge, because the majority of LV networks are data-scarce, i.e. there are no time-series phase current data throughout a year from these LV networks [86].

A number of references focus on developing phase swapping strategies for data-rich distribution networks. Reference [9] uses mixed integer programming to develop phase swapping strategies. References [16], [110] develop optimal phase swapping strategies using simulated annealing and immune algorithms, respectively. Reference [17] applies a fuzzy function to develop phase swapping strategies. Reference [111] applies an expert system to develop phase swapping strategies. References [112], [113] applied heuristic algorithms to develop phase swapping strategies, considering load patterns. Reference [114] develops phase swapping strategies using look-ahead optimizations, considering load uncertainties. Reference [48] summarizes different methods for developing phase swapping strategies. All the above references perform phase rebalancing based on full data, including network topology, time-series network current data, and demand data. However, these data are not normally available in most LV networks that are data-scarce.

Reference [115] uses smart meter data for phase swapping. However, the use of smart meters for phase balancing face three limitations: 1) a smart meter does not know which phase a customer is connected to, thus offering limited support for phase

swapping. 2) In the UK, electricity suppliers (i.e. retailers) and distribution network operators (DNOs) are separate entities. Data protection concerns arise if suppliers are to share smart meter data with DNOs. 3) In the UK, not all customers have smart meters – the rollout of smart meters is much slower than the original plan of deploying smart meters for all customers by 2020. Reference [116] uses automated meter management (AMM) system for phase balancing. The AMM system overlaps smart meters in terms of functionality: they both provide customer side data. Therefore, the AMM system faces the same limitations as smart meters, despite that the former provides additional data compared to smart meters. Further, the deployment of the AMM system for millions of LV networks in the UK is economically infeasible.

This paper advances from existing references by extrapolating the knowledge to data-scarce LV networks with neither time-series network measurements nor customer metering data. The knowledge extrapolation was an unanswered question, and it calls for a statistical approach. The extrapolation is also one of the key technical aspects of this paper. In light of this, this paper makes the following original contributions:

1) It for the first time develops phase swapping guidance for data-scarce LV networks with neither the need for any time-series network measurements nor the need for any customer-side metering data.

2) To achieve 1), this paper proposes a new statistical approach.

The statistical approach effectively overcomes the insufficient data challenge by extrapolating knowledge from a set of 800 representative data-rich LV networks with time-series phase current data to the vast population of data-scarce networks. Given the 800 data-rich LV networks, the first step is to develop a rebalancing model to rebalance the three phases of the data-rich networks by applying a non-negative matrix factorization method. The model also outputs the rebalancing potentials (the reduction

of the phase imbalance degree) of these data-rich LV networks. At the 2nd step, a rapid screening model is developed to identify the data-scarce LV networks with high rebalancing potentials among all data-scarce networks. At the 3rd step, phase swapping guidance is developed for the identified data-scarce LV networks by applying the statistical rebalancing model developed in the first step. The phase swapping guidance guide the distribution network operators to reallocate loads among the three phases of data-scarce networks in order to rebalance the three phases.

The statistical approach develops phase swapping guidance for data-scarce networks, which take the majority of the LV networks in the UK, while requiring only a minimal amount of data. The approach is economically appealing in the sense that no cost in monitoring system is incurred in Scenario 1 (where only yearly average data is required) and a minimal cost in monitoring effort is incurred in Scenario 2 (where only one-day's time-series data are required). This is compared to investing in monitoring systems to collect year-round time-series data from millions of LV networks in the UK. If distribution network operators (DNOs) follow the phase swapping guidance, energy losses would be reduced, and the network capacity that is wasted by phase imbalance would be released.

The rest of this paper is organized as follows: Chapter 5.3 presents the statistical approach. Chapter 5.4 performs case studies. Chapter 5.5 presents the detailed benefits by utilising the developed approach. Chapter 5.6 concludes this paper.

5.3. Methodology

To develop phase swapping guidance for data-scarce LV networks, this paper proposes a new statistical approach. It consists of three steps. Fig. 5-1 shows the flowchart of the statistical approach.

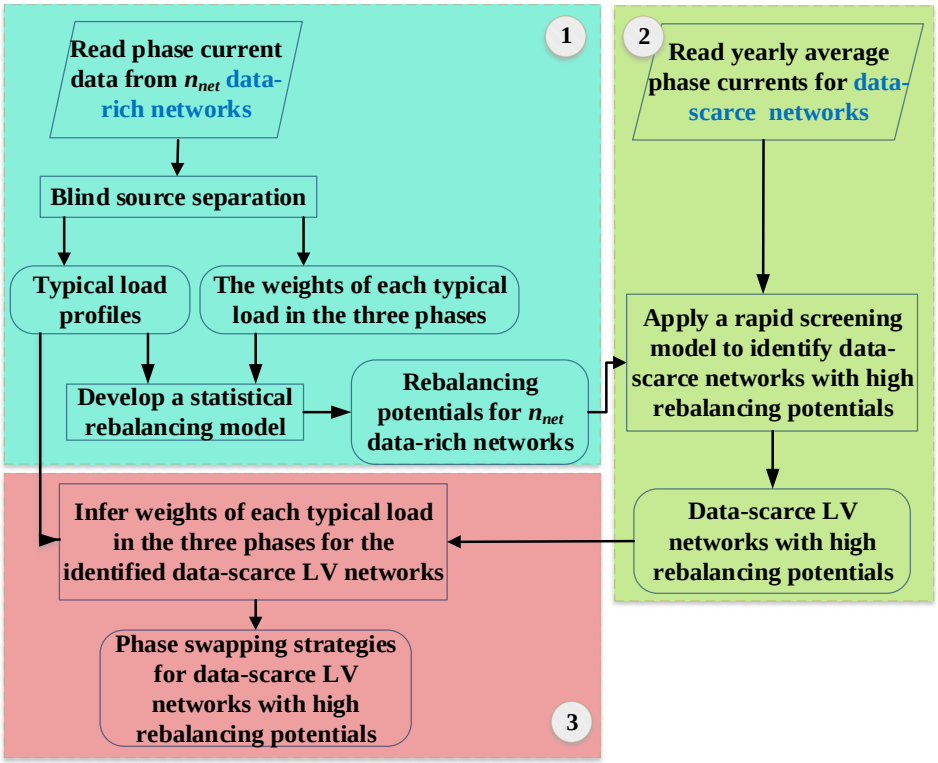


Fig. 5-1 Methodology of the statistical approach

We have the time-series phase current collected every 10 minutes throughout a year from the substations of 800 data-rich LV networks within Western Power Distribution (a UK DNO)'s business area. These LV networks cover approximately 10% of the population in South Wales areas with a good mixture of urban, suburban and rural areas and a good mixture of domestic and commercial loads [1]. These data come from the project "Low Voltage Network Template" and are described in detail in [1].

The purpose of Stage 1 is that it derives two key variables (which are not normally available in data-scarce LV networks) from data-rich networks for the development of phase swapping guidance. These two variables are 1) time-series constituent load profiles; and 2) their weights on each of the three phases. The reason for developing a rapid screening model in Stage 2 is that it identifies the data-scarce LV networks with high rebalancing potentials. These are seriously unbalanced networks that are worth

phase rebalancing. In Stage 3, the phase swapping guidance developed in Stage 1 is extrapolated from data-rich networks to data-scarce networks that have high rebalancing potentials.

5.3.1. Develop a statistical rebalancing model

For the 800 data-rich LV networks, a non-negative matrix factorization (NMF) method is adopted to extract typical load profiles and their weights in the three phases of the data-rich networks. The reason for using the NMF method is that it is a classical method to extract non-negative source signals (e.g. typical constituent load profiles in this case) and their weighting factors from mixed signals (e.g. time-series phase current data from the 800 data-rich LV networks) [105], [106]. In addition, NMF removes outliers [117]. After deriving typical constituent load profiles and their weighting factors, a statistical phase rebalancing model is developed for data-rich LV networks.

1) Extract constituent load profiles and their weights

To apply the NMF method, the time-series phase current data $\mathbf{I}_{p\emptyset,n}(t)$ from the 800 data-rich LV networks form an input phase current matrix $\mathbf{I_{PI}}$, which is given by

$$\mathbf{I_{PI}}(t) = \left[I_{p\emptyset,1}(t), I_{p\emptyset,2}(t), \cdots, I_{p\emptyset,n_{net}}(t)\right] \quad \text{6-(1)}$$

where $I_{p\emptyset,i}(t)$ denotes the time-series phase current data for phase $\emptyset$ ($\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$) of the i_{th} data-rich LV networks; $\mathbf{I}_{\mathrm{PI}}$ is a matrix with n_t (the length of the time series phase current data) rows and n_p (the total number of phases for all data-rich LV networks) columns; n_{net} is the number of data-rich LV networks. It is straightforward to see that $n_p = 3n_{net}$.

Given the input phase current matrix $\mathbf{I}_{\mathrm{PI}}$, the NMF method is applied to find the

constituent load profiles and weighting factors in each of the phases. The relationship among $\mathbf{I}_{\mathrm{PI}}$, **W**, and **H** is given by:

$$\mathbf{I}_{\mathrm{PI}} \approx \mathbf{W} \cdot \mathbf{H} \tag{6-(2)}$$

where **W**, given by 5-(3), is a matrix of *n* non-negative constituent load profiles; **H**, given by 5-(4), denotes the matrix of weighting factors. **W** and **H** are given by:

$$\mathbf{W} = \left[\boldsymbol{W_1}, \boldsymbol{W_2}, \cdots, \boldsymbol{W_{n_d}}\right] \tag{6-(3)}$$

$$\mathbf{H} = \left[H_1, H_2, \cdots, H_{n_d}\right]^{\mathrm{T}} \tag{6-(4)}$$

where W_i is the i_{th} constituent load profile; H_i is the weighting factors of the i_{th} load profile in each of the phases in $\mathbf{I}_{\mathrm{PI}}$; n_d is the number of the constituent load profiles. **W** is an n_t-by-n_d matrix. **H** is an n_d-by-n_p matrix.

The constituent loads are interpreted by three typical load profiles in the UK [110], [118]: low demand households, high demand households (the households with electric heating), and commercial loads. Therefore, in this paper, the number of the constituent load profiles n_d = 3.

To obtain **W** and **H**, an optimization model is formulated, which is the key to the NMF method. The optimization model minimizes the distance between the actual phase current $\mathbf{I}_{\mathrm{PI}}$ and the reconstructed phase current $(\mathbf{W} \cdot \mathbf{H})$**.** This optimization model is given by [105]:

$$\min_{\mathbf{W}, H_i} \sum_{i=1}^{n_p} \left\| I_{PI,i} - \mathbf{W} \cdot H_i \right\|^2 \quad \text{s.t. all elements of } \mathbf{W} \text{ and } H_i \geq \mathbf{0} \tag{6-(5)}$$

where n_p is the total number of phases for all data-rich LV networks; $I_{PI,i}$ is the i_{th} columns of the input phase current matrix $\mathbf{I}_{\mathrm{PI}}$, which was given by 5-(1); H_i is defined in 5-(4). The detailed procedure for deriving **W** and **H** is given in [105].

In this paper, the constituent load profiles derived above are normalized to be within a

band [0, 1]. The normalization is given by

$$\mathbf{W} = \left[\frac{W_1}{W_{1,max}}, \frac{W_2}{W_{2,max}}, \cdots, \frac{W_{n_d}}{W_{n_d,max}}\right] \quad 6\text{-}(6)$$

where $W_{i,max}$ denotes the maximum element of W_i, which is defined in 5-(3).

Correspondingly, the matrix **H** is adjusted, as given by

$$\mathbf{H} = \left[H_1 \cdot W_{1,max}, H_2 \cdot W_{2,max}, \cdots, H_{n_d} \cdot W_{n_d,max}\right]^{\mathrm{T}} = [h_{\emptyset,i}] \quad 6\text{-}(7)$$

where H_i and $W_{i,max}$ are defined in 5-(4) and 5-(6), respectively; $h_{\emptyset,i}$ is the normalized weighting factors in phase $\emptyset$ $(\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\})$ of the i_{th} data-rich LV network; $h_{\emptyset,i}$ is a vector of n_d rows.

2) Develop phase swapping guidance for data-rich LV networks

After deriving the constituent load profile matrix **W** and the weighting factor matrix **H**, a statistical phase rebalancing model is developed for the 800 data-rich LV networks.

For a data-rich network, a weighting factor matrix $\mathbf{H}_{net}$ is given by:

$$\mathbf{H}_{net} = [h_{\mathrm{a},}\, h_{\mathrm{b}}, h_{\mathrm{c}}] \quad 6\text{-}(8)$$

where $h_{\emptyset}$ is the weighting factors in phase $\emptyset$ $(\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\})$ of the data-rich network, as defined in 5-(7); $\mathbf{H}_{net}$ is a n_d-by-3 matrix. n_d is defined in 5-(4).

Then, a balanced weighting factor matrix $\mathbf{H}_{\mathrm{bm}}$ is derived.

$$H_b = \frac{1}{3}\sum_{i=1}^{3} H_{net,i} \quad 6\text{-}(9)$$

$$\mathbf{H}_{bm} = [H_b, H_b, H_b] \quad 6\text{-}(10)$$

where $H_{net,i}$ denotes the weighting factors at the i_{th} column (phase) of $\mathbf{H}_{net}$, which is given by 5-(8).

Secondly, a statistical phase swapping matrix (**SPSS**) is given by:

$$\mathbf{SPSS} = \left[spss_{i,j}\right] = \mathbf{H}_{bm} - \mathbf{H}_{net} \quad \text{6-(11)}$$

where $\mathrm{spss_{i,j}}$ is the i_{th} row and j_{th} column of **SPSS**.

In this paper, $spss_{i,j}$ indicates that the amount (in an average power (kW)) of the i_{th} constituent load (the i_{th} column of **W**) should be moved away from the j_{th} phase ($j = 1, 2 \text{ and } 3$ represent phase a, b and c, respectively) of the data-rich network. The rebalanced time-series phase current profiles for the data-rich LV network are given by:

$$I_{P\emptyset B} = I_{p\emptyset} - \mathbf{W} \cdot spss_i \quad \text{6-(12)}$$

where $I_{P\emptyset B}$ is a vector of n_t rows ($\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$). $I_{p\emptyset}$ is the time-series phase current data for phase $\emptyset$, as defined in 5-(1). Compared to 5-(1), the subscript *i* of $I_{p\emptyset,i}(t)$ is dropped, because now I consider a given data-rich network. **W** is defined in 5-(6). $spss_i$ is the i_{th} column of **SPSS**, which is given by 5-(11).

In addition, this model is the key to calculating the rebalancing potential (i.e. the reduction of the phase imbalance degree) for the data-rich LV network. Before deriving the rebalancing potentials, the following variables are defined.

$$I_{maxP\emptyset}(t) = \max\{I_{pa}(t), I_{pb}(t), I_{pc}(t)\} \quad \text{6-(13)}$$

$$I_{aveP\emptyset}(t) = \left(I_{pa}(t) + I_{pb}(t) + I_{pc}(t)\right)/3 \quad \text{6-(14)}$$

where $I_{pa}(t)$, $I_{pb}(t)$, and $I_{pc}(t)$ are defined in 5-(1).

$$I_{maxP\emptyset B}(t) = \max\{I_{PaB}(t), I_{PbB}(t), I_{PcB}(t)\} \quad \text{6-(15)}$$

$$I_{aveP\emptyset B}(t) = \left(I_{PaB}(t) + I_{PbB}(t) + I_{PcB}(t)\right)/3 \quad \text{6-(16)}$$

where $I_{P\emptyset B}(t)$ denotes the rebalanced time-series phase current data for phase $\emptyset$ ($\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$) of the data-rich network.

For the data-rich network, the original phase imbalance degree ($DPIB_{ori}$) and the phase imbalance degree after rebalancing ($DPIB_{Bal}$) are given by:

$$DPIB_{ori} = \frac{1}{n_t}\sum_{t=1}^{n_t}\frac{(I_{maxP\emptyset}(t) - I_{aveP\emptyset}(t))}{I_{maxP\emptyset}(t)} \quad 6\text{-}(17)$$

$$DPIB_{Bal} = \frac{1}{n_t}\sum_{t=1}^{n_t}\frac{(I_{maxP\emptyset B}(t) - I_{aveP\emptyset B}(t))}{I_{maxP\emptyset B}(t)} \quad 6\text{-}(18)$$

where $I_{maxP\emptyset}(t)$ is defined in 5-(13); $I_{aveP\emptyset}(t)$ is defined in 5-(14); $I_{maxP\emptyset B}(t)$ is defined in 5-(15); $I_{aveP\emptyset B}(t)$ is defined in 5-(16). n_t is defined in 5-(1).

Thus, the rebalancing potential RP for the data-rich LV network is given by:

$$RP = DPIB_{ori} - DPIB_{Bal} \quad 6\text{-}(19)$$

5.3.2.Develop a rapid screening model

At this stage, a rapid screening model is developed to identify data-scarce LV networks with high rebalancing potentials. Phase balancing for these networks would lead to significant benefits in terms of reducing energy losses and saving network investment costs. Before developing the rapid screening model, a virtual phase imbalance degree ($DPIB_v$) is defined:

$$DPIB_v = \frac{\max\{I_{ya}, I_{yb}, I_{yc}\} - (I_{ya}+I_{yb}+I_{yc})/3}{\max\{I_{ya}, I_{yb}, I_{yc}\}} \quad 6\text{-}(20)$$

where I_{ya}, I_{yb} and I_{yc} denote the average currents for phases a, b and c, respectively, throughout a year. $DPIB_v$ is a feature for both data-rich and data-scarce networks.

The reason for using $DPIB_v$ as the feature is that: 1) it is derived from yearly average phase current data, which are available for both data-rich and data-scarce LV networks [87]; 2) the alternative yearly maximum phase current data from data-scarce LV networks cannot represent the actual phase imbalance [2].

Based on the virtual phase imbalance degree and the rebalancing potentials from n_{net} data-rich LV networks, a rapid screening model is developed. It uses a quadratic

function to map the virtual phase imbalance degree ($DPIB_v$) to the rebalancing potential, which is given by

$$RP = f(DPIB_v) = \theta_1 DPIB_v{}^2 + \theta_2 DPIB_v + \theta_3 \quad \text{6-(21)}$$

where RP is defined in 5-(19). $DPIB_v$ is the virtual phase imbalance degree; θ_1, θ_2 and θ_3 are coefficients.

The choice of a quadratic function is justified by Fig. 5-4 (in the case study section), which shows an approximate quadratic relationship between the degree of phase imbalance and the rebalancing potential. In other words, the quadratic function represents an optimal trade-off between bias and variance. The fitted quadratic function can then be used to estimate the rebalancing potentials for data-scarce LV networks.

To derive the optimal coefficients (θ_1, θ_2 and θ_3), the following optimization model is solved:

$$\min_{\theta_1,\ \theta_2,\ \theta_3} \sqrt{\frac{1}{n_{net}} \sum_{i=1}^{n_{net}} (RP_i - f(DPIB_{v,i}))^2} \quad \text{6-(22)}$$

where $f(DPIB_{v,i}) = \theta_1 DPIB_{v,i}{}^2 + \theta_2 DPIB_{v,i} + \theta_3$

$DPIB_{v,i}$ is the virtual phase imbalance degree for the i_{th} data-rich LV networks; RP_i denotes the derived rebalancing potentials for the i_{th} data-rich LV networks; n_{net} is the number of data-rich LV networks.

After developing the rapid screen model, rebalancing potential is estimated for data-scarce LV networks. Furthermore, a threshold variable RP_T divide rebalancing potentials into two sections: the low rebalancing potential section (LR) and the high rebalancing potential section (HR). Thus, a data scarce LV network is identified as a network with high rebalancing potential, if the estimated rebalancing potential is greater than RP_T. The choice of RP_T (i.e. how “high” is a high rebalancing potential) is

subjective. It requires expert's judgment, and the criteria can vary from case to case. For example, RP_T can be chosen so that the "high rebalancing potential" section includes the LV networks whose rebalancing potentials are among the top 25% of all LV networks considered.

5.3.3. Develop phase swapping guidance for data-scarce LV networks with high rebalancing potentials

In this section, phase swapping guidance is developed for data-scarce LV networks with high rebalancing potentials (defined in Chapter 5.3.2). The weights of each constituent load are rebalanced in the three phases according to the developed guidance, thus approximately balancing the three phases. The key is to infer the weighting factor matrix **H** for data-scarce LV networks. When applied to the field, different distribution network operators have different types of available data: 1) one scenario is where only the yearly average phase current data (I_{ya}, I_{yb}, I_{yc}) are available; 2) the other scenario is where only one day's phase current data $\mathrm{I}_{o\emptyset}(t)$ $(\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\})$ are available. The above two scenarios are considered.

1) Use the yearly average phase current data to infer the missing weighting factors

Select a data-scarce network as an example, the weighting factor matrix $\mathbf{H}_{\mathrm{ds}}$ is inferred, as given by:

$$H_{ds,i} = \left[\frac{\frac{1}{3}I_{ya}}{w_{ave.i}}, \frac{\frac{1}{3}I_{yb}}{w_{ave.i}}, \frac{\frac{1}{3}I_{yc}}{w_{ave.i}}\right] \quad \text{6-(23)}$$

$$\mathbf{H}_{\mathrm{ds}} = \begin{bmatrix} H_{ds,1} \\ H_{ds,2} \\ \vdots \\ H_{ds,n_d} \end{bmatrix} \quad \text{6-(24)}$$

where $w_{ave.i}$ is the average value of the data in the i_{th} column of **W** (given by 5-(6));

n_d is defined in 5-(3).

After deriving the weighting factor matrix $\mathbf{H}_{\mathrm{ds}}$, a balanced weighting factor matrix $\mathbf{H}_{\mathrm{bms}}$ is derived for the data-scarce network, which is given by:

$$H_{bs} = \frac{1}{3}\sum_{i=1}^{3} \mathbf{H}_{ds,i} \quad 6\text{-}(25)$$

$$\mathbf{H}_{bms} = [H_{bs}, H_{bs}, H_{bs}] \quad 6\text{-}(26)$$

where $\mathbf{H}_{ds,i}$ denotes the weighting factors at the i_{th} column of $\mathbf{H}_{\mathrm{ds}}$.

Finally, for the data-scarce LV network, a statistical phase swapping matrix ($\mathbf{SPSS_s}$) is given by:

$$\mathbf{SPSS_s} = \left[spss_{s,i,j}\right] = \mathbf{H}_{bms} - \mathbf{H}_{ds} \quad 6\text{-}(27)$$

where $\mathbf{H}_{bms}$ and $\mathbf{H}_{ds}$ are derived in 5-(26) and 5-(24), respectively; where $spss_{s,i,j}$ is the i_{th} row and j_{th} column of $\mathbf{SPSS_s}$. The variable $\mathrm{spss}_{s,i,j}$ indicates that the amount of the i_{th} constituent load (the i_{th}column of **W**, as defined in 5-(6)) should be moved away from the j_{th} phase ($j = 1, 2 \text{ and } 3$ represent phases a, b and c, respectively) of the data-scarce network.

2) Use one day's phase current data to infer the missing weighting factors

Select a data-scarce network as an example, the weighting factor matrix $\mathbf{H}_{\mathrm{ds}}$, as defined in 5-(24), is inferred through the following steps.

Firstly, for the i_{th} constituent load (the i_{th} column of **W**), the j_{th} day's load profile $W_{i,j}$ is derived, which is given by

$$W_{i,j} = \mathbf{W}((n_{dt}(j-1)+1):(n_{dt}j), i) \quad 6\text{-}(28)$$

where $\mathbf{W}((n_{dt}(j-1)+1):(n_{dt}j), i)$ is the i_{th} column, $(n_{dt}(j-1)+1)$ to $(n_{dt}j)$ rows of the matrix **W,** which is define in 5-(6). **W** (*x*, *y*) denotes the element in the x_{th} row and

y_{th} column of matrix **W**. The "*m*:*n*" expression denotes "from *m* to *n* (inclusive)", e.g. from row $(n_{dt}(j-1)+1)$ to row $(n_{dt}j)$ inclusive of both rows. $W_{i,j}$ is a vector of n_{dt} rows; n_{dt} is the length of one-day's time-series phase current data.

Then, an average load profile that corresponds to the i_{th} constituent load is given by:

$$I_{pd,i} = \frac{1}{n_{dy}} \sum_{j=1}^{n_{dy}} W_{i,j} \tag{6-(29)}$$

$I_{pd,i}$ is a vector of n_{dt} rows; n_{dt} is defined in 5-(28); $W_{i,j}$ is defined in 5-(28); n_{dy} is the number of days throughout a year.

To derive the weighting factor matrix for a data-scarce network, an optimization problem is solved:

$$\begin{gathered} \min_{h_{1,\emptyset},h_{2,\emptyset},\cdots,h_{n_d,\emptyset}} \sqrt{\sum_{t=1}^{n_{dt}} (I_{o\emptyset}(t) - \mathbf{W}_{\mathrm{ds}}(t) \cdot H_{ds,\emptyset})^2} \\ s.t.\ h_{1,\emptyset}, h_{2,\emptyset}, \cdots and\ h_{n\_d,\emptyset} \geq 0 \\ \text{where } \mathbf{W}_{\mathrm{ds}} = [I_{pd,1}, I_{pd,2}, \cdots I_{pd,n_d}]; \\ H_{ds,\emptyset} = [h_{1,\emptyset}, h_{2,\emptyset}, \cdots, h_{n_d,\emptyset}]^T \end{gathered} \tag{6-(30)}$$

$I_{pd,i}$ is define in 5-(29);$h_{i,\emptyset}$ denotes the weight of the i_{th} constituent load in phase $\emptyset$ ($\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$) of the data-scarce network; $I_{o\emptyset}(t)$ is the t_{th} element of one day's phase current from the data-scarce network; $\mathbf{W}_{\mathrm{ds}}$ is a matrix with n_{dt} rows and n_d columns. $\mathbf{W}_{\mathrm{ds}}(t)$ denotes the elements in the t_{th} row of matrix $\mathbf{W}_{\mathrm{ds}}$; n_{dt} is defined in 5-(28).

The weighting factor matrix for the data-scarce LV network $\mathbf{H}_{\mathrm{ds}}$ is given by:

$$\mathbf{H}_{\mathrm{ds}} = [H_{ds,a}, H_{ds,b}, H_{ds,c}] \tag{6-(31)}$$

where $H_{ds,\emptyset}$ ($\emptyset \epsilon \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$) is defined in 5-(30).

After deriving $\mathbf{H}_{\mathrm{ds}}$, the statistical phase swapping matrix ($\mathbf{SPSS}_s$) is given by 5-(25) – 5-

(27). $\mathbf{SPSS}_s$ presents the phase swapping matrix. $\mathbf{SPSS}_s$ has the same meaning as that explained immediately after Equation 5-(27).

5.3.4.Method for validation

Before validation, the rebalancing potential RP for n_{net} ($n_{net} = 800$) data rich LV networks are derived in stage 1 (explained in Chapter 5.3.1) as the accurate RP.

Then, to validate the developed phase swapping guidance for data-scarce LV networks, *k*-fold cross-validation ($k = 10$ in this paper) is used. Firstly, the 800 data-rich LV networks are randomly partitioned into k equal-sized groups. Then, the LV networks from one of the k groups are held out as the validation samples (treat them as if there were data-scarce LV networks), the LV networks from the remaining $k - 1$ groups are used as the training samples (data-rich LV networks). The training samples are used to develop the statistical rebalancing model and rapid screening model as explained in Chapter 5.3.1 and 5.3.2. Then, the phase swapping guidance is developed for the validation samples.

For the first scenario (explained in Chapter 5.3.3), the rebalanced time-series phase current data $\mathbf{I}_{SB}$ for a validation sample is given by:

$$\mathbf{I}_{SB} = [I_{pa}, I_{pb}, I_{pc}] - \mathbf{W} \cdot \mathbf{SPSS}_s \quad \text{6-(32)}$$

where **W** is defined in 5-(6); $\mathbf{SPSS}_s$ is defined in 5-(27); $I_{p\emptyset}$ is defined in 5-(1) ($\emptyset \epsilon \{a, b, c\}$); Compared to 5-(1), the subscript *i* of $I_{p\emptyset,i}(t)$ is dropped, because now I consider a given network.

The corresponding phase imbalance degree after rebalancing is given by:

$$DPIB_{sba} = \frac{1}{n_t} \sum_{t=1}^{n_t} \frac{(I_{maxSB}(t) - I_{aveSB}(t))}{I_{maxSB}(t)} \quad \text{6-(33)}$$

where $I_{maxSB}(t)$ is the maximum element in the t_{th} row of $\mathbf{I}_{SB}$, which is defined in 5-(32); $I_{aveSB}(t)$ is the average value of the elements in the t_{th} row of $\mathbf{I}_{SB}$, which is defined in 5-(32); n_t is defined in 5-(1).

The validation rebalancing potential for the validation sample is given by:

$$RP_v = DPIB_{ori} - DPIB_{sba} \quad \text{6-(34)}$$

where $DPIB_{ori}$ denotes the original phase imbalance degree for the validation sample (given by 5-(17)); $DPIB_{sba}$ denotes the phase imbalance degree after rebalancing for the validation sample (given by 5-(33)).

Through the previous validation, the rebalancing error for the i_{th} validation sample e_v is given by

$$e_v = \frac{RP - RP_v}{RP} \quad \text{6-(35)}$$

where RP (given by 5-(19)) is the accurate rebalancing potential for the validation sample; RP_v is the rebalancing potential (given by the proposed statistical rebalancing method) for the validation sample. This error indicates the validity of the phase swapping performance for the data-scarce LV networks.

The previous process repeats until each of the *k* groups has been held out as the validation samples. After *k* iterations, the rebalancing potentials (i.e. the reduction of phase imbalance degree) are derived for all data-scarce LV networks. The rebalancing errors are given by 5-(35).

For the second scenario (explained in Chapter 5.3.3), the validation process in each iteration of the *k*-folds validation is presented in Fig. 5-2.

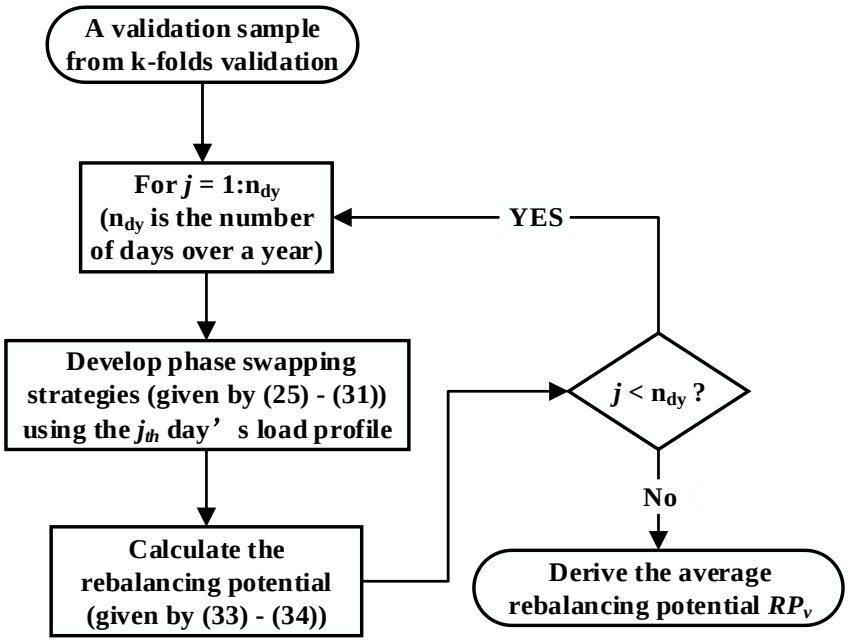


Fig. 5-2 Flowchart of the validation process for Scenario 2)

After deriving RP_v for the validation sample, the rebalancing error is given by 5-(35). The above process repeats until each of the *k* groups has been held out as the validation samples.

5.4. Case Studies

This section presents the numerical results. The results from the statistical rebalancing modelling and rapid screening modelling for data-rich networks are given in Chapters 5.4.1 and 5.4.2, respectively. Chapter 5.4.3 presents the phase swapping results for data-scarce networks. A discussion is presented in Chapter 5.4.4.

5.4.1. Results from the statistical rebalancing model

In the first step, three constituent load profiles are extracted and scaled to the range of [0,1] by the maximum value throughout a year. These three constituent loads are low demand households, high demand households (customers with electric heating, shown

as in Figure 253 of reference [118]) and commercial loads. Fig. 5-3 presents the three constituent load profiles throughout a week.

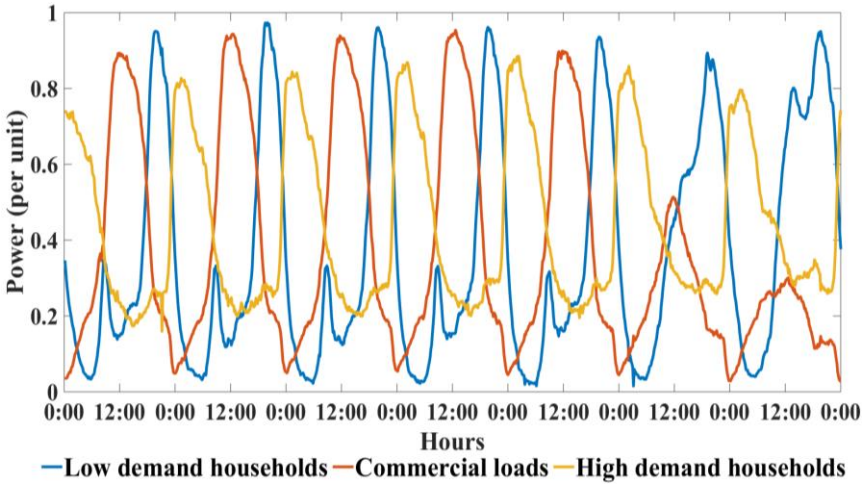


Fig. 5-3 The constituent load profiles throughout an example week (Monday to Sunday)

Commercial loads and low demand households share the same characteristic: the profile is different between workdays and weekends. For commercial loads, the weekend peak load is approximately 1/3 of the workday peak. For low demand households, the first peak of the weekend's load is approximately twice of the first peak of the workday's load.

Select a data-rich network as an example. If the average power of these typical customers (e.g. low demand household, high demand household and commercial load) are 0.2 kW, 0.4 kW and 0.8 kW, respectively, the phase swapping guidance are presented as follows:

TABLE 5-1

A STATISTICAL PHASE SWAPPING GUIDANCE

Unit: kW, number of loads

	Phase a	Phase b	Phase c
Low demand households	–5.01, –25	+4.43, +22	+0.57, +3
High demand households	–2.61, –7	+3.10, +8	–0.49, -1
Commercial loads	–2.90, –4	+0.62, +1	+2.28, +3

In TABLE 5-1, a negative number indicates the amount of the constituent load that should be moved away to other phases; a positive number indicates the amount of the constituent load that should be taken in from other phases. For example, for phase a, 25 low demand households (which sums up to an average power of 5.01 kW), 7 high demand households (which sums up to an average power of 2.61 kW), and 4 commercial loads (which sums up an average power of 2.90 kW) should be moved away to other phases. If phase swapping strictly follows the guidance in TABLE 5-1, the amount of each constituent load in the three phases is rebalanced, thus balancing the three phases. For example, after phase swapping, the average power of low demand household is 15 kW in phases a, b and c, respectively, but it was 10.8 kW, 20.28 kW and 16.42 kW before phase swapping. The accurate rebalancing potential (*RP*) is 0.115.

If the derived phase swapping guidance is followed, it significantly reduces the degree of phase imbalance for the 800 data-rich LV networks. It reduces the average and maximum phase imbalance degree by 34% and 40%, respectively. In addition, 387 (48.3%) data-rich LV networks have rebalancing potentials greater than 0.05. Five (0.67%) out of the 800 data-rich LV networks have negative rebalancing potentials, indicating that phase swapping actually increases the degree of phase imbalance. The reason for this is that these few LV networks have no consistent direction of imbalance

among the three phases. In other words, phase swapping is not applicable to these five networks.

5.4.2.Results from the rapid screening model

Fig. 5-4 shows a quadratic mapping from the virtual degree of phase imbalance to the rebalancing potential, based on the data from the 800 data-rich LV networks. Such a mapping is the rapid screening model.

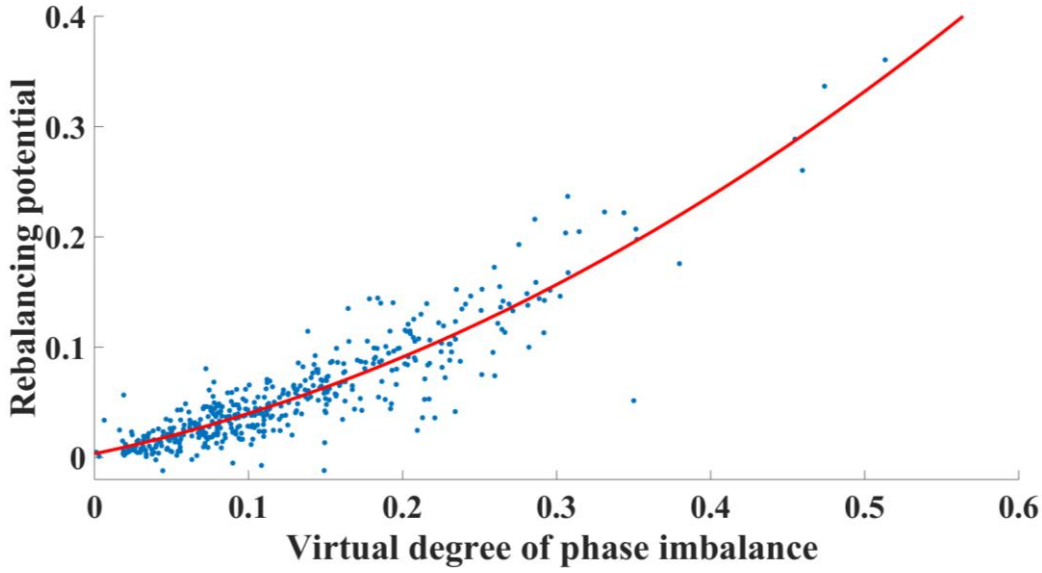


Fig. 5-4 The rapid screening model

The majority of the networks have $DPIB_v$ values that fall in the range of [0, 0.3] and rebalancing potentials that fall in the range of [0, 0.2]. In this paper, the threshold of the rebalancing potential is $RP_T = 0.05$. 48.3% of the LV networks have rebalancing potentials greater than the threshold – they have high rebalancing potentials.

Based on the data set of the 800 networks, the mapping incurs a root-mean-squared error (RMSE) and a mean absolute percentage error (MAPE) of 0.0214 and 16.3%, respectively.

5.4.3.Phase swapping guidance for data-scarce LV networks with high rebalancing potentials

To rebalance the data-scarce LV networks with high rebalancing potentials (explained in Chapter 5.4.2), two scenarios are considered: 1) the scenario with yearly average phase current data; and 2) the scenario with one day's phase current data. Phase swapping guidance are developed for the two scenarios. The results are validated by 10-fold cross-validation.

1) Phase swapping results for data-scarce networks with yearly average phase current

Select the data-rich network (the example in Chapter 5.4.1) as a validation sample (treat this data-rich network as if it were data-scarce by ignoring its time-series data), this network has yearly average phase currents $[I_{ya}, I_{yb}, I_{yc}] = [90.43\text{A}, 168.48\text{A}, 144.63\text{A}]$. The weighting factors matrix is calculated as follows:

$$\mathbf{H}_{\text{ds}} = \begin{bmatrix} 81.74 & 159.27 & 108.36 \\ 81.74 & 159.27 & 108.36 \\ 81.74 & 159.27 & 108.36 \end{bmatrix}$$

If the average power of these typical customers (e.g. low demand household, high demand household and commercial load) are 0.2 kW, 0.4 kW and 0.8 kW, respectively, the phase swapping guidance are presented as follows:

TABLE 5-2

A STATISTICAL PHASE SWAPPING GUIDANCE

Unit: kW, number of loads

	Phase a	Phase b	Phase c
Low demand households	–3.52, –17	+2.85, +14	+0.67, +3
High demand households	–3.37, –9	+2.72, +7	+0.65, +2
Commercial loads	–4.26, –5	+3.43, +4	+0.81, +1

The meaning of negative numbers and positive numbers are explained after TABLE 5-1. For example, for phase b, 14 low demand households (which sums up to an average power of 2.85 kW), 7 high demand households (which sums up to an average power of 2.72 kW), and 4 commercial loads (which sums up an average power of 3.43 kW) should be taken in from other phases. The rebalancing potential for this network is 0.0959. The rebalancing error e_v (defined in 5-(35)) is 16%. It indicates that, for this validation sample, the rebalancing potential RP_v (given by the proposed method) is 16% lower than the accurate *RP*.

Through validation, the average rebalancing error is 19.33% in Scenario 1). If the phase swapping implementation strictly follows the developed guidance, the practical benefits (including network reinforcement cost reduction [77] and energy loss reduction [73]) from phase swapping are shown as follows:

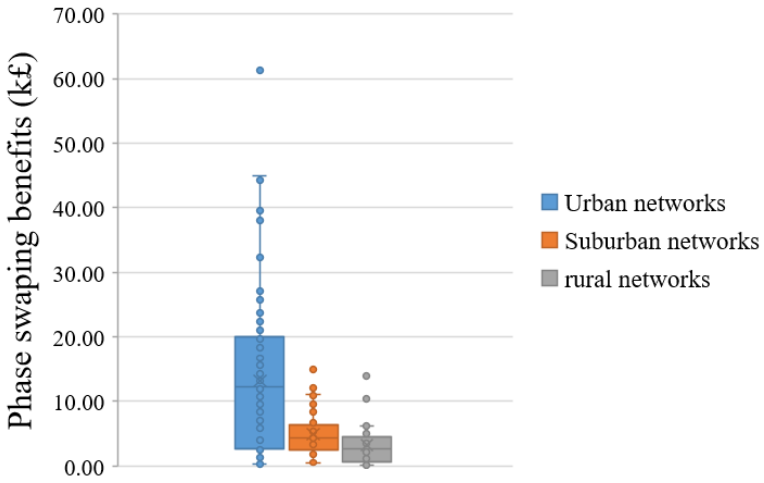


Fig. 5-5 Practical benefits form phase swapping

In Fig. 5-5, the average and maximum phase swapping benefits are: 1) £12,232 and £61,304, respectively, for urban LV networks; 2) £4,940 and £20,372, respectively, for suburban LV networks; and 3) £3,280 and £13,919, respectively, for rural LV networks.

2) Use one day's phase current data to infer the missing weighting factors

In this scenario, two average load profiles (given by 5-(25)) are considered: 1) average workday profile; 2) average weekend profile. The average one day's load profiles (scaled to the range of [0,1] by the maximum value throughout a year) are shown as follows:

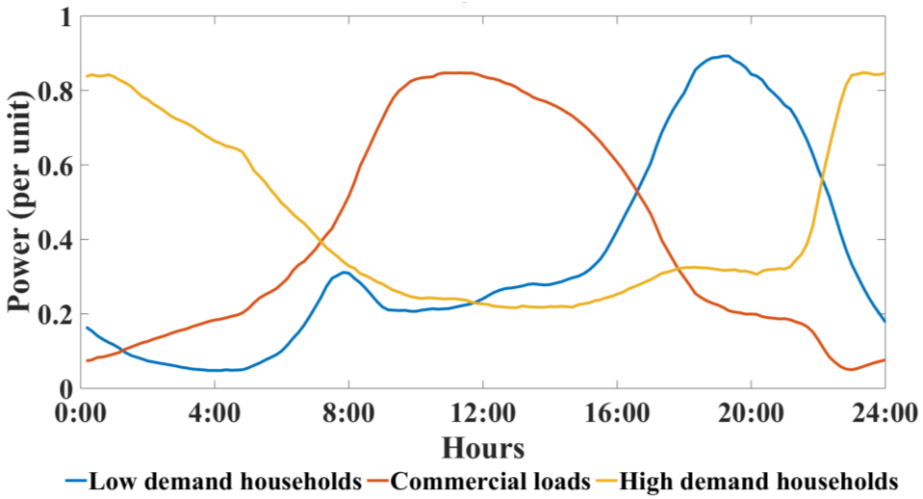


Fig. 5-6 Average workday profiles of constituent loads

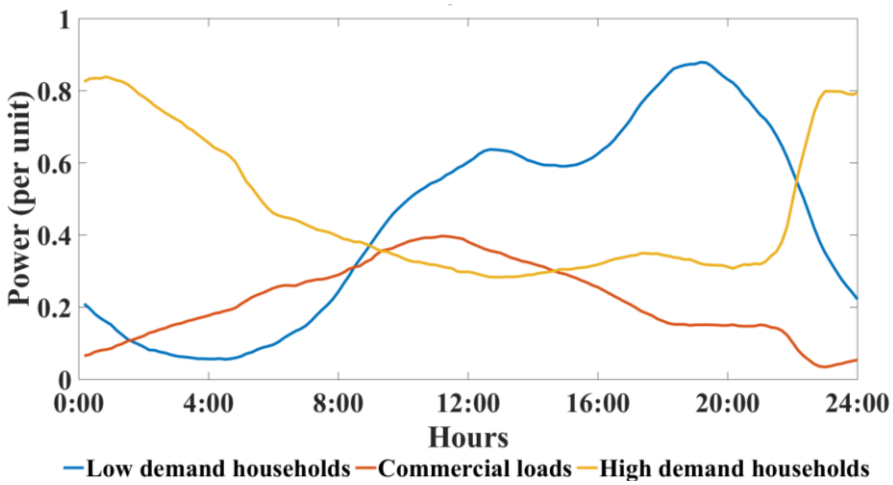


Fig. 5-7 Average weekend profiles of constituent loads

Select the data-rich network (the example in Chapter 5.4.1) as a validation sample (treat this data-rich network as if it were data-scarce), this network has one workday's phase current data. If the average power of these typical customers (e.g., low demand household, high demand household and commercial load) are 0.2 kW, 0.4 kW and 0.8

kW, respectively, the phase swapping guidance are presented as follows:

TABLE 5-3

A STATISTICAL PHASE SWAPPING GUIDANCE

Unit: kW, number of loads

	Phase a (kW)	Phase b (kW)	Phase c (kW)
Low demand households	–6.90, –35	+6.14, +31	+0.76, +4
High demand households	–3.34, -8	+4.04, +10	–0.70, –2
Commercial loads	–2.12, –3	–0.35. –1	+2.47, +4

The meaning of negative numbers and positive numbers are explained after TABLE 5-1. For example, for phase a, 35 low demand households (which sums up to an average power of 6.90 kW), 8 high demand households (which sums up to an average power of 3.34 kW), and 3 commercial loads (which sums up an average power of 2.12 kW) should be moved away to other phases. The rebalancing potential for this validation sample is 0.1037. The rebalancing error e_s (defined in 5-(35)) is approximately 10.1%. It indicates that, for this validation sample, the rebalancing potential RP_v (given by the proposed method) is 10.1% lower than the accurate *RP*.

Through validation, the average rebalancing errors are: 1) 14.3%, using the workday's phase current data; 2) 31.9%, using the weekend's phase current data for the data-scarce networks. In this case, using the workday's phase current data as the feature for the data-scarce network results in a greater reduction of the phase imbalance degree compared with using the weekend's phase current data. If the phase swapping implementation strictly follows the developed guidance, the practical benefits (including network reinforcement cost reduction [77] and energy loss reduction [73]) from phase swapping are shown as follows:

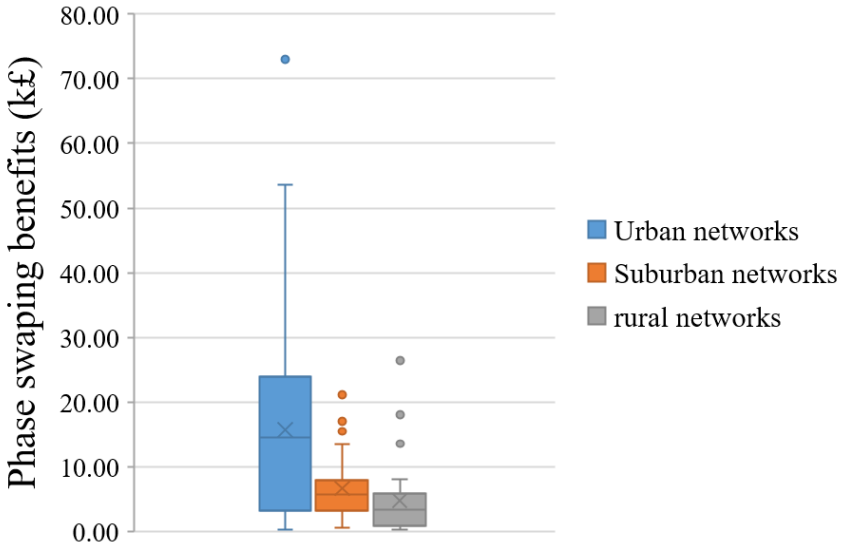


Fig. 5-8 Practical benefits form phase swapping

In Fig. 5-8, the average and maximum phase swapping benefits are: 1) £15,884 and £72,981, respectively, for urban LV networks; 2) £6,674 and £21,125, respectively, for suburban LV networks; and 3) £4,710 and £26,457, respectively, for rural LV networks.

The average rebalancing errors of 19.33% and 14.3% are acceptable because my statistical approach requires minimal data from data-scarce networks: in Scenario 1), only yearly average phase currents are required; in Scenario 2), only one day's time-series phase currents are required. There is a trade-off between the data requirement from data-scarce networks and the accuracy (measured by the rebalancing error) of the phase swapping guidance derived by the statistical approach. The more data required from these networks, the lower the rebalancing error will be obtained, but more costs will be needed to collect the additional data. In addition, the practical benefits presented in Fig. 5-5 and Fig. 5-8 demonstrate the effectiveness of the statistical approach.

5.4.4.Implementation

The developed approach is not a full phase balancing strategy but rather provides

important guidance for phase balancing for data-scarce LV networks at a minimal cost of monitoring. The guidance does not specify which connection point on the feeder is to be phase swapped and how, as this depends on the specific LV feeder topology and customers' phase connectivity which vary from case to case. Rather, the approach provides the following key guidance for phase swapping:

1) whether any given data-scarce network suffers from a serious phase imbalance or not;

2) whether any given data-scarce network has a phase imbalance direction or not. A phase imbalance direction refers to the existence of a particular phase that is consistently heavier (or lighter) than the other phases. Phase swapping is only applicable where there is a phase imbalance direction.

3) given any data-scarce network, move what load profiles from which phase to which phase in order to achieve near-balanced three phases (the results are presented in TABLE 5-2 and TABLE 5-3).

In other words, the above 1) and 2) inform whether any given LV network is worthy of phase swapping or not. If yes, to develop a network-specific phase swapping strategy, the following steps should be taken:

1) The DNOs should first obtain the network topology.

2) The type of each customer should be identified. The customers' phase connectivity should also be obtained.

3) Determine the points of phase swapping so that the phase swapping strategy closely follows the 3rd guidance as mentioned above.

5.4.5.Discussions

The developed statistical approach addresses a problem that no existing method can address: developing phase swapping guidance for data-scarce LV networks with neither network monitoring nor any metering from the customer side. The phase swapping guidance derived through the statistical approach serves as a benchmark. Future research can compare the effectiveness of their guidance with the one in this paper.

The statistical approach is designed to be generic. To apply this method to other countries, it requires the following steps: 1) Collect yearly time-series phase current data from *N* number of LV networks (*N* should be at least 800 hundred). These LV networks should be representative enough. 2) Set the number of constituent loads, so that the constituent loads are interpretable in that country, e.g. the low demand households in the UK. Then, the proposed method can be applied.

It should be noted that there are millions of LV networks in the UK alone. The fact that the statistical approach only requires the training data from 800 representative networks is not demanding, compared to requiring full data from each of the millions of networks in the UK. Furthermore, distribution network operators (DNOs) can reasonably monitor the full data for 800 networks at a moderate cost.

Phase balancing is most needed at LV (11kV/415V) substations. This is because a substation is a critical load node seen by higher-level networks. Phase balancing at the substation would prevent phase imbalance and its consequences from propagating to higher-level networks. Depending on individual circumstances, phase balancing may be extended beyond substations onto critical nodes on LV feeders. However, phase balancing at every node of the LV network is neither necessary nor feasible. Therefore, phase swapping is not required at all connection points but are only required at critical

nodes, e.g. the substation. This significantly relieves the burden of phase balancing on a mass scale. Further, not all LV substations need phase balancing, only those with serious phase imbalance need balancing, thus further relieving the burden of phase balancing.

This paper uses the average load profiles to approximate customers' loads. This approximation is justified in the following way: 1) the phase swapping guidance derived by the statistical approach, which uses the average load profiles, turns out to have satisfactory accuracy. This implicitly justifies the use of the average load profiles. 2) The derived load profiles are typical load profiles for different types of customers (e.g. low demand households, high demand households and commercial loads) because the non-negative matrix factorization has clustering property.

Training data of less durations are also used to develop the phase swapping guidance. The accuracies of the guidance are presented in TABLE 5-4:

TABLE 5-4

EFFECTIVENESS COMPARISON OF DEVELOPED PHASE SWAPPING GUIDANCE

	Year-round data	Half-year data	One month data	One week data
Rebalancing error	14.3%	16.9%	22.11%	27.4%

The above results prove that my current practice of using year-round data yields the least rebalancing error. A greater rebalancing error indicates a lower reduction of phase imbalance (hence less effectiveness of the phase swapping guidance) for the data-scarce networks.

The statistical approach does not divide the training data from the 800 LV networks into urban, suburban, and rural groups, because: 1) If the division were made, the training data in each group would be insufficient, thus compromising the accuracy of the developed phase swapping guidance; and 2) the case studies yield phase swapping

guidance of satisfactory accuracies for data-scarce LV networks. This in turn justifies the practice of training the model on the 800 LV networks as a whole rather than dividing these networks into three groups. Although the average phase current values are not yet collected from all LV networks, they can be obtained via the following means at minimal costs:1) The average phase current values can be derived from energy meter data if the energy consumption is recorded per phase. 2) The average phase current values can be obtained from the protection systems, which monitor the network operation status over time. 3) A recent project, OpenLV, sponsored by Western Power Distribution and undertaken by EA Technology, monitors a range of LV (11kV/415V) substations and the collected data include the average phase current values [119]. This paper advocates the collection of the average phase current values for the purpose of developing phase swapping guidance.

The developed approach yields effective phase swapping guidance with satisfactory accuracy for the sample networks I have. These networks have a low penetration of PVs and EVs, representing the status quo in Western Power Distribution's business areas. To account for increasing single-phase PVs and EVs, the approach can be adapted by updating the average load profiles to account for single-phase PVs and EVs. This also requires the monitoring of representative PV/EV-rich, data-rich substations. Then the developed approach can learn the knowledge and extrapolate it to PV- or EV-rich, data-scarce substations. This is part of the future work.

5.5. The detailed benefits by utilizing the developed phase swapping guidance

In Chapter 5.4 presents the theoretical benefits from the phase swapping guidance by the developed statistical approach. This section discusses the benefits in detail. The calculation steps are presented as follows:

- Given the phase swapping results of the examples in Fig. 5-5 and Fig. 5-8, the pre-balancing three-phase load profiles and post-balancing three-phase load profiles are obtained.

- Taken equations 2-(1) and 2-(3) and the pre-balancing three-phase load profiles and post-balancing three-phase load profiles of the example LV networks calculate the reduced energy losses and reduced reinforcement costs. Fig. 5-5 and Fig. 5-8 provides the total reduced costs if the utility price is 0.18 £/kWh and the following network reinforcement costs [12]:

TABLE 5-5

PARAMETERS AND REINFORCEMENT COSTS FOR MAIN FEEDERS

Asset	Area	Circuit Length (km)	Investment Cost per Unit (£/km)
Underground cable	Urban	0.2	67200
Underground cable	Suburban	0.3	16400
Overhead line	Rural	0.4	15000

TABLE 5-6

PARAMETERS AND REINFORCEMENT COSTS FOR TRANSFORMERS

Area	Transformer Capacity (kVA)	Investment Cost (£)
Urban	400	26400
Suburban	259	16100
Rural	150	5800

It should be stressed that urban networks have 5 feeders connected to the substation. Suburban and rural networks have 3.5 and 1.5 feeders, respectively, connected to the substation[120].

- If data-scarce LV networks only have the yearly average phase currents, the detailed reduced energy losses and reduced reinforcement costs for urban, suburban and rural networks are shown as follows:

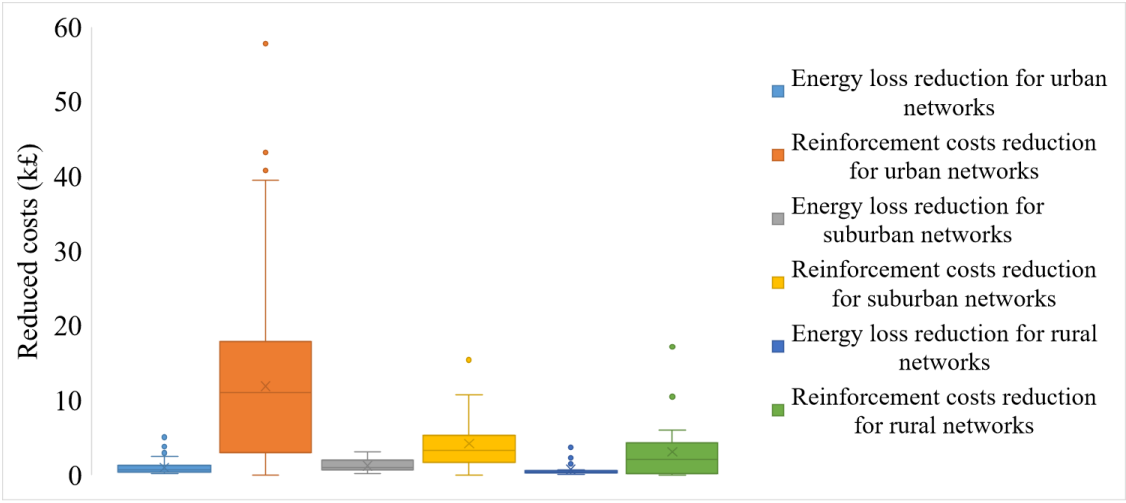


Fig. 5-9 The detailed benefits for urban, suburban and rural data-scarce LV networks with only the yearly average phase currents

- If data-scarce LV networks have one workday's time-series phase current data, the detailed reduced energy losses and reduced reinforcement costs for urban, suburban and rural networks are shown as follows:

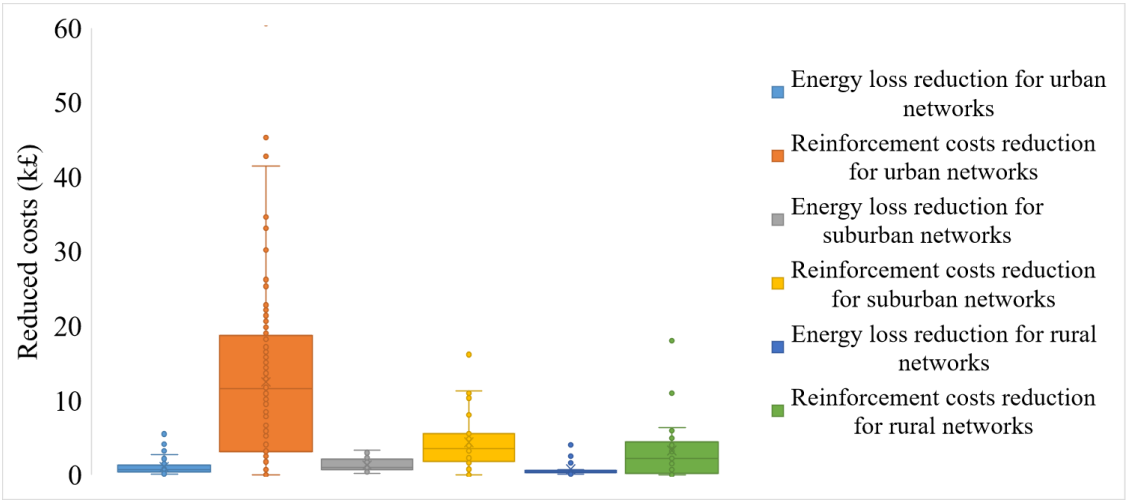


Fig. 5-10 The detailed benefits for urban, suburban and rural data-scarce LV networks with one workday's phase current data

5.6. Conclusions

This paper addresses an unresolved problem for distribution network operators (DNOs): develop phase swapping guidance for data-scarce low voltage (415V, LV) networks with neither time-series network measurements nor customer metering data. To achieve this, this paper develops a new statistical phase swapping approach, extrapolating knowledge from 800 representative data-rich networks to data-scarce networks. This approach produces phase swapping guidance that guides the DNOs to reallocate typical loads (e.g. low demand households, high demand households and commercial loads) among the three phases, thus rebalancing the three phases of data-scarce networks.

Case studies are performed to validate the statistical phase swapping approach, which achieves effective reductions of the phase imbalance degrees for data-scarce networks: the reduction of phase imbalance degree is only 14.3% lower than that for data-rich networks. If DNOs follow the phase swapping guidance produced by the statistical approach, energy losses would be reduced and the network capacity wasted by phase imbalance would be released, while only a minimal amount of data is required.

5.7. Chapter summary

This chapter, for the first time, develops a statistical approach to make phase swapping guidance for data-scarce LV networks. Case studies reveal that the statistical approach produces effective phase swapping guidance, which reduces the phase imbalance degrees for 99% of the data-scarce LV networks, and the maximum reduction is 0.35. Moreover, the reduction of phase imbalance degree for data-scarce LV networks is only 14.3% lower than that for data-rich networks.

It should be noted that this chapter does not develop comprehensive phase swapping

strategies, where the network topologies and smart meter data of customers are required. Given that there are over 900,000 LV networks in the UK, my approach is more generic and data-efficient than comprehensive phase swapping strategies. In turn, the generic advantage implies my approach cannot achieve maximum phase balancing effectiveness as a comprehensive phase swapping strategy can achieve.

This chapter also discovers two findings:

- In discovering the rebalancing potential for data-scarce LV networks, I find that not all LV networks can be rebalanced by phase swapping, particularly the LV networks naturally with low virtual phase imbalance degree (calculated from the yearly average phase currents by equation 5-(20)). This implies that, for these LV networks, its load composition is much more complex than imagined. Further investigation of the load composition is required in future works.

- In developing phase swapping guidance for data-scarce LV networks, if DNOs could collect one day's substation-side load profiles, one weekday's data prevails than one weekend's data. The case studies in Chapter 5.4.3 justify that using one weekday's data to developing phase swapping guidance for data-scarce LV networks doubles the phase swapping effectiveness compared to using one weekend's data.

Chapter 6.

Conclusions and Future Works

Chapter contents:

This chapter presents thesis conclusions and three potential future works.

6.1. Conclusions

Phase imbalance is a widespread and long-outstanding problem for LV networks. A number of references have developed phase imbalance assessment methods and phase balancing methods. However, in reality, the data-scarcity problem limits the implementation of existing phase imbalance diagnosis methods and phase balancing methods, as the majority of LV networks do not collect time-series phase current data due to the lack of advanced monitoring devices. To get around this limitation, this thesis, for the first time, develops three approaches.

- Phase imbalance raises energy losses on the three phases for LV networks. This thesis develops a regression-based approach to estimate imbalance-induced phase energy losses for data-scarce LV networks. This approach is data-efficient and highly scalable as it only requires yearly average and maximum phase current data. Case studies reveal that: for 90% of the data-scarce LV networks in urban, suburban and rural areas, the average estimation accuracies are 80.6%, 88.2% and 87.8%, respectively.

- Phase imbalance causes the appearance of phase residual currents, which flow from customers to the transformer's neutral point via the neutral wire or the groud. The phase residual current flow causes additional energy losses. This thesis develops a range-estimation-based approach to estimate imbalance-induced residual energy losses for data-scarce LV networks. This approach only requires yearly average phase currents for data-scarce LV networks, thus being more data-efficient than the aforementioned regression-based approaches. Moreover, this approach derives a confidence range of imbalance-induced residual energy losses to promote estimation credibility in accommodating future phase imbalance changes. Through validation, this approach delivers correct range estimation for over 82% of data-scarce LV networks.

- Phase swapping is one of the classical ways to address phase imbalance. This thesis originally develops a statistical approach to make phase swapping guidance to promote the practicality of phase swapping for massive implementation. Compared to existing references, the statistical approach does not require year-round substation-side phase current data and customer-side smart meter data, thus applying for data-scarce LV networks. Case studies reveal that the statistical approach produces effective phase swapping guidance, which reduces the phase imbalance degrees for 99% of the data-scarce LV networks, and the maximum reduction is 0.35. Moreover, the reduction of phase imbalance degree for data-scarce LV networks is only 14.3% lower than that for data-rich networks.

The above three approaches addressed unsolved problems in the industry– assessing imbalance-induced energy losses and making phase swapping guidance for data-scarce LV networks. Compared to existing approaches, the developed approaches have advantages in terms of data-efficient, low costs and scalability. Further, the first two approaches 1) turn decision making for phase balancing investment into possibility in the industry; and 2) visualising the connection between imbalance-induced costs and non-operation data, thus could support improving LV network planning for DNOs.

However, similar to all data-driven approaches, the developed approaches also have limitations in the implementation.

- The estimation accuracy for estimating imbalance-induced energy losses and the effectiveness for developing phase swapping guidance are sensitive to input data. This thesis suggests that DNOs should select representative data-rich LV networks as the input to improve estimation accuracy. For example, this thesis uses time-series phase current data collected from 800 data-rich LV substations throughout a year at an interval of 15 minutes. These substations, within Western Power Distribution (a UK DNO)'s business area, cover a good mix of geographical areas

(urban, suburban, and rural) and customer composition (domestic, commercial, and industrial).

- This thesis combines different machine learning methods together to customise the aforementioned statistical approaches. For example, the machine learning method is robust-linear regression in Chapter 3.3. However, there is no guarantee that the determined robust-linear regression fits all DNOs' datasets. In the implementation, DNOs should adopt other classic methods and use the developed evaluation methods by this thesis to determine which machine learning method is applicable for their dataset. For example, the classic alternative methods for robust-linear regression are regression tree, SVR, or Gaussian processing model.

- In the implementation, all of the developed statistical approaches should be updated at least every year to promote their accuracy and effectiveness.

Last but not least, this thesis discovers two findings that might only exist for LV networks:

- In assessing imbalance-induced energy losses, the developed statistical approaches deliver significantly high estimation errors for LV networks with drastic low imbalance-induced energy losses. The error can achieve up to 800%. However, these outliers show very low imbalance-induced energy losses, which are only up to 0.3MWh ((which costs £54 additional losses if the electricity price is £0.18/kWh)). Therefore, before training the developed approach, it should remove the LV networks with significant lower APELs from the training data. This prevents the tuned parameters of the trained model from the impact of the outlier LV networks, thus improving the estimation accuracy for the majority of non-outlier LV networks.

- In discovering the rebalancing potential for data-scarce LV networks, I find that not all LV networks can be rebalanced by phase swapping, particularly the LV

networks naturally with low virtual phase imbalance degree (calculated from the yearly average phase currents by equation 5-(20)). This implies that for these LV networks, its load composition is much more complex than imagined. Further investigation is required in future works in terms of detailed load composition.

- In developing phase swapping guidance for data-scarce LV networks, if DNOs could collect one day's substation-side load profiles, one weekday's data prevails than one weekend's data. The case studies in Chapter 5.4 justify that using one weekday's data to developing phase swapping guidance for data-scarce LV networks doubles the phase swapping effectiveness compared to using one weekend's data.

6.2. Future works

Previous chapters have reviewed existing literature and originally addressed the data-scarce problem in assessing imbalance-induced energy losses and making phase swapping strategies for LV networks. In this chapter, I propose three potential future works that have not been studied in this thesis. Chapter 5.1.1 discusses how to derive a credible assessment of imbalance-induced capacity wastes for data-scarce LV networks. Chapter 5.1.2 discusses the implication of load distribution changes on estimating energy losses for data-scarce LV networks, especially the implication in estimating imbalance-induced energy losses. Chapter 5.1.3 discusses a new incentive scheme to encourage flexible customers to provide phase balancing.

6.2.1. Estimating imbalance-induced capacity wastes for data-scarce LV networks

Chapter 2.2.2 notes that phase imbalance incurs capacity wastes which result in additional reinforcement costs. Reference [12] and [15] developed formulas to quantify

imbalance-induced reinforcement costs. Reference [45] converted reinforcement costs calculation for individual network into the utility-scale. However, previous studies do not consider that peak load growth in LV networks is significantly uncertain [121], [122]. Using one year's peak load data to calculate LV networks' additional reinforcement costs is not credible in making long-term phase balancing investment decisions. Moreover, for the majority of LV networks, three-phase total peak data are not recorded due to lacking advanced monitoring devices. It, therefore, raises a research question: how to deliver credible imbalance-induced capacity waste assessments for data-scarce LV networks.

To address the problem, I originally propose a probability-based statistical approach:

- This approach will deliver a probability study of imbalance-induced additional reinforcement costs instead of a point data study delivered by existing literature. For each LV network, the probability study does not require data from multiple years. It, therefore, is data-efficient and cost-efficient for massive deployments.

- This approach would be straightforward and engineering-friendly.

6.2.2. The implication of load distribution changes on estimating imbalance-induced energy losses for data-scarce LV networks

In previous studies, energy losses on LV networks are calculated by three means: 1) using exact network topology and time-series data to calculate energy losses; 2) using load loss factor to estimate energy losses; and 3) using time-series phase current data and load distribution assumptions to estimate energy losses. Given that the majority of LV networks are data-scarce, i.e. they do not collect time-series phase current, this thesis has developed two statistical approaches to estimate imbalance-induced energy

losses for data-scarce LV networks, assuming that loads are always triangle or rectangle distributed along the LV feeders [123]. These two distributions are commonly used to estimate energy losses. However, with the growing connection of LCT devices, which are principally heavy loads, the load distribution would vary from time to time. Given that the start-up time of heat pumps and EV charging are random for each customer, the load distribution will perform significant uncertainty. An example is performed in Fig. 6-1.

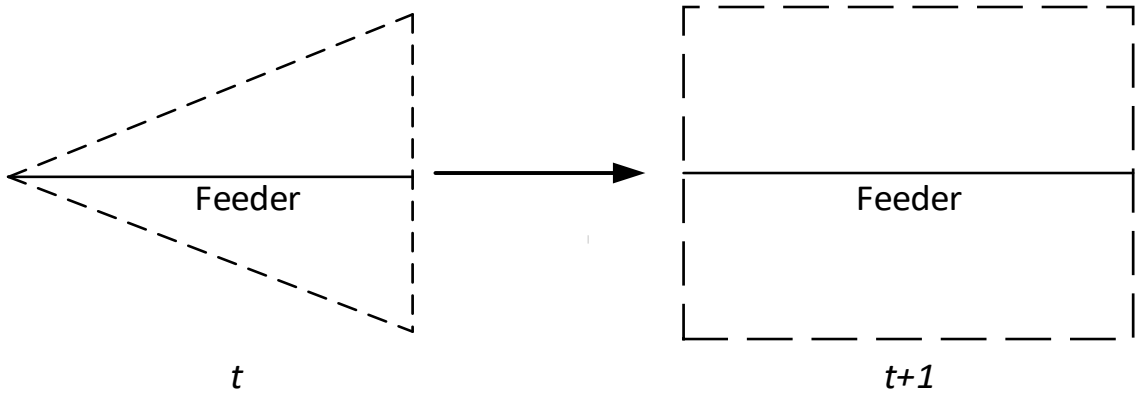


Fig. 6-1 An example of load distribution change on the time horizon

Fig. 6-1 present that at time t, the load are triangle distributed along the LV feeder. However, at time t+1, the load distribution changes to the rectangle. For triangle distribution, the power loss is estimated by [123]:

$$P = \frac{8}{15} I^2 R \qquad 6\text{-}(1)$$

For rectangle distribution, the power loss is estimated by [123]:

$$P = \frac{1}{3} I^2 R \qquad 6\text{-}(2)$$

For LV networks with the bulk of LCT loads, the load distribution would not constantly be triangle or rectangle on time horizons. For the example case in Fig. 6-1, assuming loads are always triangle distributed along the feeder would overestimate the energy losses. By contrast, assuming loads are always rectangle distributed along the feeder

would underestimate the energy losses. Combining the varying load distribution problem with the data scarcity problem for LV networks, therefore, raises a challenge: what is the average load distribution on time horizons for future data-scarce LV networks and how to estimate imbalance-induced energy losses.

6.2.3. An incentive scheme to encourage flexible customers for phase balancing

Chapter 2.3 and 1.1.2 have discussed existing phase balancing solutions and their pros and cons. However, the scalability and costs for existing approaches still require investigation, especially when Ofgem had reduced the rate of turn to 3.9% – 4.2% in 2020. In light of the above problems, references [20] and [21] developed a converter-based technical solution called three-phase converter dispatching. This solution centrally reallocates the phase currents of grid-connected three-phase AC/DC converters to rebalance the three phases at the substation side of distribution networks. Each AC/DC converter is intentionally controlled to operate in an unbalanced mode through an advanced control logic [22], [23] to achieve phase balancing at the substation side of LV networks. In reality, a growing number of customers are supplied via three phases [24] and have grid-connected three-phase converters such as three-phase EV charging poles, three-phase DC heat pumps, DC micro-grids, three-phase energy storage systems, three-phase distributed PVs, and wind turbines. Furthermore, reference [25] used the flexibility inherent in single-phase EVs to deliver phase balancing.

Therefore, it has the potential to engage both three-phase and single-phase flexible customers for phase balancing. However, a gap remains between the technical control algorithm and real business implementation: there is currently no incentive scheme that could motivate sufficient flexible customers to prioritize the predominant consequence

of phase imbalance for each LV network.

To this end, I originally propose an incentive scheme to bridge this gap. This incentive scheme is designed to be practical and economically feasible under the UK's regulatory mechanism. This scheme will consider both competition impacts [124] and peer impacts [125], [126] to promote inclusiveness for flexible customers of all sizes.

References

[1] "LV network templates for a low-carbon future," 2014, https://www.westernpower.co.uk/docs/Innovation/Closed-projects/Network-Templates/LVNT-Appendix-A-Knowledge-Management.aspx.

[2] "HV and LV Phase Imbalance - SP Energy Networks," 2015, https://www.spenergynetworks.co.uk/userfiles/file/HVandLVPhaseImbalanceAssessment16.pdf.

[3] J. Y. Yong, V. K. Ramachandaramurthy, K. M. Tan, and N. Mithulananthan, “A review on the state-of-the-art technologies of electric vehicle, its impacts and prospects,” *Renewable and Sustainable Energy Reviews,* vol. 49, pp. 365-385, 2015/09/01/, 2015.

[4] *Distribution Future Energy Scenarios*, UKPN, 2021.

[5] P. Lico, M. Marinelli, K. Knezović, and S. Grillo, “Phase balancing by means of electric vehicles single-phase connection shifting in a low voltage Danish grid,” in 2015 50th International Universities Power Engineering Conference (UPEC), 2015, pp. 1-5.

[6] T. Routtenberg, Y. Xie, R. M. Willett, and L. Tong, “PMU-Based Detection of Imbalance in Three-Phase Power Systems,” *IEEE Transactions on Power Systems,* vol. 30, no. 4, pp. 1966-1976, 2015.

[7] J. D. Watson, N. R. Watson, and I. Lestas, “Optimized Dispatch of Energy Storage Systems in Unbalanced Distribution Networks,” *IEEE Transactions on Sustainable Energy,* vol. 9, no. 2, pp. 639-650, 2018.

[8] L. F. Ochoa, R. M. Ciric, A. Padilha-Feltrin, and G. P. Harrison, “Evaluation of distribution system losses due to load unbalance,” in 15th Power Systems Computation Conference PSCC 2005, 2005.

[9] Z. Jinxiang, C. Mo-Yuen, and Z. Fan, “Phase balancing using mixed-integer programming [distribution feeders],” *IEEE Transactions on Power Systems,* vol. 13, no. 4, pp. 1487-1492, 1998.

[10] D. Singh, R. K. Misra, and S. Mishra, “Distribution system feeder re-phasing

considering voltage-dependency of loads," *International Journal of Electrical Power & Energy Systems,* vol. 76, pp. 107-119, 2016.

[11] R. G. Harley, E. B. Makram, and E. G. Duran, "The effects of unbalanced networks and unbalanced faults on induction motor transient stability," *IEEE Transactions on Energy Conversion,* vol. 3, no. 2, pp. 398-403, 1988.

[12] K. Ma, R. Li, and F. Li, "Quantification of Additional Asset Reinforcement Cost From 3-Phase Imbalance " *IEEE Transactions on Power Systems,* vol. 31, no. 4, pp. 2885 - 2891 July, 2016.

[13] W. H. Kersting, "The computation of neutral and dirt currents and power losses," in IEEE PES Power Systems Conference and Exposition, 2004., 2004, pp. 213-218 vol.1.

[14] S. Pajic, and A. E. Emanuel, "Effect of Neutral Path Power Losses on the Apparent Power Definitions: A Preliminary Study," *IEEE Transactions on Power Delivery,* vol. 24, no. 2, pp. 517-523, 2009.

[15] K. Ma, R. Li, and I. Hernando-Gil, "Quantification of Additional Reinforcement Cost From Severe Three-Phase Imbalance," *IEEE Transactions on Power Systems,* vol. 32, no. 5, pp. 4143-4144, Sept. , 2017.

[16] J. Zhu, G. Bilbro, and C. Mo-Yuen, "Phase balancing using simulated annealing," *IEEE Transactions on Power Systems,* vol. 14, no. 4, pp. 1508-1513, 1999.

[17] R. A. Hooshmand, and S. Soltani, "Fuzzy Optimal Phase Balancing of Radial and Meshed Distribution Networks Using BF-PSO Algorithm," *IEEE Transactions on Power Systems,* vol. 27, no. 1, pp. 47-57, 2012.

[18] S. H. Soltani, M. Rashidinejad, and A. Abdollahi, "Dynamic phase balancing in the smart distribution networks," *International Journal of Electrical Power & Energy Systems,* vol. 93, pp. 374-383, 2017/12/01/, 2017.

[19] S. U. Haq, B. Arif, A. Khan, and J. Ahmed, "Automatic three phase load balancing system by using fast switching relay in three phase distribution system," in 2018 1st International Conference on Power, Energy and Smart Grid (ICPESG), 2018, pp. 1-6.

[20] S. Liu, X. Cui, Z. Lin, Z. Lian, Z. Lin, F. Wen, Y. Ding, Q. Wang, L. Yang, R. Jin, and H. Qiu, "Practical Method for Mitigating Three-Phase Unbalance Based on

Data-Driven User Phase Identification," *IEEE Transactions on Power Systems,* vol. 35, no. 2, pp. 1653-1656, 2020.

[21] T. Hong, and F. d. León, "Centralized Unbalanced Dispatch of Smart Distribution DC Microgrid Systems," *IEEE Transactions on Smart Grid,* vol. 9, no. 4, pp. 2852-2861, 2018.

[22] S. Weckx, and J. Driesen, "Load Balancing With EV Chargers and PV Inverters in Unbalanced Distribution Grids," *IEEE Transactions on Sustainable Energy,* vol. 6, no. 2, pp. 635-643, 2015.

[23] L. San-Yi, and W. Chi-Jui, "On-line reactive power compensation schemes for unbalanced three phase four wire distribution feeders," *IEEE Transactions on Power Delivery,* vol. 8, no. 4, pp. 1958-1965, 1993.

[24] V. B. Bhavaraju, and P. N. Enjeti, "Analysis and design of an active power filter for balancing unbalanced loads," *IEEE Transactions on Power Electronics,* vol. 8, no. 4, pp. 640-647, 1993.

[25] S. Chen, Z. Guo, Z. Yang, Y. Xu, and R. S. Cheng, "A Game Theoretic Approach to Phase Balancing by Plug-in Electric Vehicles in the Smart Grid," *IEEE Transactions on Power Systems*, pp. 1-1, 2019.

[26] S. Sun, B. Liang, M. Dong, and J. A. Taylor, "Phase Balancing Using Energy Storage in Power Grids Under Uncertainty," *IEEE Transactions on Power Systems,* vol. 31, no. 5, pp. 3891-3903, 2016.

[27] V. Rigoni, L. F. Ochoa, G. Chicco, A. Navarro-Espinosa, and T. Gozel, "Representative Residential LV Feeders: A Case Study for the North West of England," *IEEE Transactions on Power Systems,* vol. 31, no. 1, pp. 348-360, 2016.

[28] J. Zhang, Y. Wang, Y. Weng, and N. Zhang, "Topology Identification and Line Parameter Estimation for Non-PMU Distribution Network: A Numerical Method," *IEEE Transactions on Smart Grid,* vol. 11, no. 5, pp. 4440-4453, 2020.

[29] "Families and households in the UK: 2020," 2 March, 2021; www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2020.

[30] M. Chindris, A. Cziker, A. Miron, H. Balan, and A. Sudria, "Propagation of unbalance in electric power systems," in 2007 9th International Conference on

Electrical Power Quality and Utilisation, 2007, pp. 1-5.

[31] S. Weckx, C. Gonzalez, and J. Driesen, “Reducing grid losses and voltage unbalance with PV inverters,” in 2014 IEEE PES General Meeting | Conference & Exposition, 2014, pp. 1-5.

[32] W. Jiye, Z. Nan, and H. Hanyong, “Three-phase imbalance prediction: A hazard-based method,” in 2016 IEEE International Conference on Power and Renewable Energy (ICPRE), 2016, pp. 226-231.

[33] "NEMA Standards Publication no. MG 1-1993," Motors and Generators.

[34] A. v. Jouanne, and B. Banerjee, “Assessment of voltage unbalance,” *IEEE Transactions on Power Delivery,* vol. 16, no. 4, pp. 782-790, 2001.

[35] "Eliminate voltage unbalance," 2000; https://www.nrel.gov/docs/fy00osti/27832.pdf.

[36] R. F. Woll, “Effect of Unbalanced Voltage on the Operation of Polyphase Induction Motors,” *IEEE Transactions on Industry Applications,* vol. IA-11, no. 1, pp. 38-42, 1975.

[37] W. H. Kersting, and W. H. Phillips, “Phase frame analysis of the effects of voltage unbalance on induction machines,” *IEEE Transactions on Industry Applications,* vol. 33, no. 2, pp. 415-420, 1997.

[38] A. A.Sallam, and O.P.Malik, *Electric Distribution System*, p.^pp. 88-95: John Wiley & Sons, Inc, 2011.

[39] J. C. Montano, P. Salmeron, and J. P. Thomas, “Analysis of power losses for instantaneous compensation of three-phase four-wire systems,” *IEEE Transactions on Power Electronics,* vol. 20, no. 4, pp. 901-907, 2005.

[40] O. K. Ignatius, A. K. Saadu, and O. S. Emmanuel, “Analysis of copper losses due to unbalanced load in a transformer,” *IJRRAS,* vol. 23, pp. 46-53, 2015.

[41] T.-H. Chen, "Evaluation of line loss under load unbalance using the complex unbalance factor," *IEE Proceedings - Generation, Transmission and Distribution,* 142, 1995, https://digital-library.theiet.org/content/journals/10.1049/ip-gtd_19951708, 1995].

[42] P. S. N. Rao, and R. Deekshit, “Energy loss estimation in distribution feeders,”

IEEE Transactions on Power Delivery, vol. 21, no. 3, pp. 1092-1100, 2006.

[43] G. Carpinelli, F. Gagliardi, A. Losi, and V. Mangoni, “A simplified method for evaluating the losses in an unbalanced three-phase power system,” *International Journal of Electrical Power & Energy Systems,* vol. 10, no. 2, pp. 117-122, 1988/04/01/, 1988.

[44] K. Malmedal, and P. K. Sen, “A Better Understanding of Load and Loss Factors,” in 2008 IEEE Industry Applications Society Annual Meeting, 2008, pp. 1-6.

[45] K. Ma, R. Li, and F. Li, “Utility-Scale Estimation of Additional Reinforcement Cost From Three-Phase Imbalance Considering Thermal Constraints,” *IEEE Transactions on Power Systems,* vol. 32, no. 5, pp. 3912-3923, 2017.

[46] M. W. Siti, D. V. Nicolae, A. A. Jimoh, and A. Ukil, “Reconfiguration and Load Balancing in the LV and MV Distribution Networks for Optimal Performance,” *IEEE Transactions on Power Delivery,* vol. 22, no. 4, pp. 2534-2540, 2007.

[47] W. M. Siti, A. Jimoh, and D. Nicolae, “Distribution network phase load balancing as a combinatorial optimization problem using fuzzy logic and Newton–Raphson,” *Electric Power Systems Research,* vol. 81, no. 5, pp. 1079-1087, 2011/05/01/, 2011.

[48] K. Wang, S. Skiena, and T. G. Robertazzi, “Phase balancing algorithms,” *Electric Power Systems Research,* vol. 96, pp. 218-224, 2013/03/01/, 2013.

[49] A. Garcés, J. C. Castaño, and M. A. Rios, "Phase Balancing in Power Distribution Grids: A Genetic Algorithm with a Group-Based Codification," *Handbook of Optimization in Electric Power Distribution Systems*, M. Resener, S. Rebennack, P. M. Pardalos and S. Haffner, eds., pp. 325-342, Cham: Springer International Publishing, 2020.

[50] S. Mansani, R. Y. Udaykumar, Santoshkumar, M. A. Asha Rani, and S. Sreejith, “Phase Balancing of DG-Integrated Smart Secondary Distribution Network,” Singapore, 2021, pp. 321-333.

[51] F. Ding, and K. A. Loparo, “Feeder Reconfiguration for Unbalanced Distribution Systems With Distributed Generation: A Hierarchical Decentralized Approach,” *IEEE Transactions on Power Systems,* vol. 31, no. 2, pp. 1633-1642, 2016.

[52] G. Schweickardt, J. M. G. Alvarez, and C. Casanova, “Metaheuristics approaches to solve combinatorial optimization problems in distribution power

systems. An application to Phase Balancing in low voltage three-phase networks," *International Journal of Electrical Power & Energy Systems,* vol. 76, pp. 1-10, 2016/03/01/, 2016.

[53] S. Yan, S. Tan, C. Lee, B. Chaudhuri, and S. Y. R. Hui, "Electric Springs for Reducing Power Imbalance in Three-Phase Power Systems," *IEEE Transactions on Power Electronics,* vol. 30, no. 7, pp. 3601-3609, 2015.

[54] S. M. Fazeli, H. W. Ping, N. B. A. Rahim, and B. T. Ooi, "Individual-phase decoupled P–Q control of three-phase voltage source converter," *IET Generation, Transmission & Distribution,* vol. 7, no. 11, pp. 1219-1228, 2013.

[55] X. J. Zeng, H. F. Zhai, M. X. Wang, M. Yang, and M. Q. Wang, "A system optimization method for mitigating three-phase imbalance in distribution network," *International Journal of Electrical Power & Energy Systems,* vol. 113, pp. 618-633, 2019/12/01/, 2019.

[56] P. Gangwar, S. N. Singh, and S. Chakrabarti, "An Analytical Approach for Phase Balancing Considering Customer Load Profile," in 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), 2019, pp. 1-5.

[57] S. Taghipour Boroujeni, M. Mardaneh, and Z. Hashemi, "A Dynamic and Heuristic Phase Balancing Method for LV Feeders," *Applied Computational Intelligence and Soft Computing,* vol. 2016, pp. 6928080, 2016/05/31, 2016.

[58] S. Yongsug, and T. A. Lipo, "Control scheme in hybrid synchronous stationary frame for PWM AC/DC converter under generalized unbalanced operating conditions," *IEEE Transactions on Industry Applications,* vol. 42, no. 3, pp. 825-835, 2006.

[59] T. Hong, and F. d. León, "Controlling Non-Synchronous Microgrids for Load Balancing of Radial Distribution Systems," *IEEE Transactions on Smart Grid,* vol. 8, no. 6, pp. 2608-2616, 2017.

[60] W. Wang, and N. Yu, "Phase Balancing in Power Distribution Network with Data Center," *SIGMETRICS Perform. Eval. Rev.,* vol. 45, no. 2, pp. 64–69, 2017.

[61] F. Nejabatkhah, and Y. W. Li, "Flexible Unbalanced Compensation of Three-Phase Distribution System Using Single-Phase Distributed Generation Inverters," *IEEE Transactions on Smart Grid,* vol. 10, no. 2, pp. 1845-1857, 2019.

[62] S. R. Gunn. "Support Vector Machines for Classification and Regression," 1997, http://m.svms.org/tutorials/Gunn1997.pdf.

[63] R. Andersen, *Modern Methods for Robust Regression* SAGE Publishing, 2007.

[64] G. A. F. Seber, and A. J. Lee, *Linear Regression Analysis*, 2rd edition ed., p.^pp. 2-4: Wiley Series in Probability and statistics, 2012.

[65] E. Schulz, M. Speekenbrink, and A. Krause, “A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions,” *Journal of Mathematical Psychology,* vol. 85, pp. 1-16, 2018/08/01/, 2018.

[66] J. D. Rodriguez, A. Perez, and J. A. Lozano, “Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 3, pp. 569-575, 2010.

[67] L. Fang, and K. Ma, “Assessment of additional phase energy losses caused by phase imbalance for data-scarce LV networks,” *IET Generation, Transmission & Distribution,* vol. 14, no. 4, pp. 675-681, 2020.

[68] P. M. S. Carvalho, L. A. F. M. Ferreira, J. J. E. Santana, A. M. F. Dias, and J. A. C. Machado, “Combined Effects of Load Variability and Phase Imbalance Onto Simulated LV Losses,” *IEEE Transactions on Power Systems,* vol. 33, no. 6, pp. 7031-7041, 2018.

[69] D. I. H. Sun, S. Abe, R. R. Shoults, M. S. Chen, P. Eichenberger, and D. Farris, “Calculation of Energy Losses in a Distribution System,” *IEEE Transactions on Power Apparatus and Systems,* vol. PAS-99, no. 4, pp. 1347-1356, 1980.

[70] L. Fang, K. Ma, R. Li, Z. Wang, and H. Shi, “A Statistical Approach to Estimate Imbalance-Induced Energy Losses for Data-Scarce Low Voltage Networks,” *IEEE Transactions on Power Systems,* vol. 34, no. 4, pp. 2825-2835, 2019.

[71] L. Jiang, J. Meng, Z. Yin, Y. Dong, and J. Zhang, "Research on additional loss of line and transformer in low voltage distribution network under the disturbance of power quality." pp. 364-369.

[72] A. L. Shenkman, “Energy loss computation by using statistical techniques,” *IEEE Transactions on Power Delivery,* vol. 5, no. 1, pp. 254-258, 1990.

[73] W. H. Kersting, *Distribution System Modeling and Analysis*, 4th Edition ed., p.^pp. 39-77: CRC Press, taylor& Francis Group, 2017.

[74] M. T. Bina, and A. Kashefi, “Three-phase unbalance of distribution systems: Complementary analysis and experimental case study,” *International Journal of Electrical Power & Energy Systems,* vol. 33, no. 4, pp. 817-826, 2011/05/01/, 2011.

[75] C. Yu, W. Yao, and X. Bai. "Robust Linear Regression: A Review and Comparison," 2014, https://arxiv.org/pdf/1404.6274v1.pdf.

[76] R. Andersen, *Modern Approaches for Robust Regression*: SAGE Publishing, 2007.

[77] K. Ma, R. Li, and F. Li, “Quantification of Additional Asset Reinforcement Cost From 3-Phase Imbalance,” *IEEE Transactions on Power Systems,* vol. 31, no. 4, pp. 2885-2891, 2016.

[78] M. Mohanpurkar, and S. Suryanarayanan, “Regression Modeling for Accommodating Unscheduled Flows in Electric Grids,” *IEEE Transactions on Power Systems,* vol. 29, no. 5, pp. 2569-2570, 2014.

[79] J. Zhang, C. Y. Chung, and Y. Han, “Online Damping Ratio Prediction Using Locally Weighted Linear Regression,” *IEEE Transactions on Power Systems,* vol. 31, no. 3, pp. 1954-1962, 2016.

[80] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, “Time-series clustering – A decade review,” *Information Systems,* vol. 53, no. Supplement C, pp. 16-38, 2015.

[81] N. Mehra, and S. Gupta, “Survey on multiclass classification methods,” *International Journal of Computer Science and Information Technologies,* vol. 4, pp. 572-576, 2013.

[82] J. G. Saw, C. K. Y. Mark, and T. C. Mo, “Chebyshev Inequality with Estimated Mean and Variance,” *The American Statistician,* vol. 38, no. 2, pp. 130-132, 1984.

[83] K. Ma, F. Li, and R. Aggarwal, “Quantification of Additional Reinforcement Cost Driven by Voltage Constraint Under Three-Phase Imbalance,” *IEEE Transactions on Power Systems,* vol. 31, no. 6, pp. 5126-5134, 2016.

[84] A. J. Urquhart, and M. Thomson, “Impacts of Demand Data Time Resolution on Estimates of Distribution System Energy Losses,” *IEEE Transactions on Power Systems,* vol. 30, no. 3, pp. 1483-1491, 2015.

[85] J. C. López, J. F. Franco, and M. J. Rider, “Optimisation-based switch allocation to improve energy losses and service restoration in radial electrical distribution systems,” *IET Generation, Transmission & Distribution,* vol. 10, no. 11, pp. 2792-2801, 2016/08/01, 2016.

[86] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, “Development of Low Voltage Network Templates—Part I: Substation Clustering and Classification,” *IEEE Transactions on Power Systems,* vol. 30, no. 6, pp. 3036-3044, 2015.

[87] "Sepam™ Series 20 Protective Relays User’s Manual," https://www.schneider-electric.com/resources/sites/SCHNEIDER_ELECTRIC/content/live/FAQS/221000/FA221290/en_US/63230-216-208C1_Sepam_Series_20_User_Manual.pdf.

[88] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*: CRC Press, 1986.

[89] S. S. Keerthi, and C. J. Lin, “Asymptotic behaviors of support vector machines with Gaussian kernel,” *Neural Comput,* vol. 15, no. 7, pp. 1667-89, Jul, 2003.

[90] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, “Time-series clustering – A decade review,” *Information Systems,* vol. 53, pp. 16-38, 2015/10/01/, 2015.

[91] B. Mirkin, *Clustering for Data Mining A Data Recovery Approach*: Chapman and Hall/CRC, 2005.

[92] S. DeDeo, R. X. D. Hawkins, S. Klingenstein, and T. Hitchcock, “Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems,” *Entropy,* vol. 15, no. 6, 2013.

[93] A. ng. "CS229 lecture notes: Support Vector Machines," 2017, http://cs229.stanford.edu/notes/cs229-notes3.pdf.

[94] J. Huang, Z. Jiang, L. Rylands, and M. Negnevitsky, “SVM-based PQ disturbance recognition system,” *IET Generation, Transmission & Distribution,* vol. 12, no. 2, pp. 328-334, 2018.

[95] Y. Freund, and R. E. Schapire, “Experiments with a new boosting algorithm,” in The Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 1966.

[96] L. Wang, D. Wang, and C. Hao, “Intelligent CFAR Detector Based on Support

Vector Machine," *IEEE Access,* vol. 5, pp. 26965-26972, 2017.

[97] Q. Leming, P. S. Routh, and K. Kyungduk, "Wavelet deconvolution in a periodic setting using cross-validation," *IEEE Signal Processing Letters,* vol. 13, no. 4, pp. 232-235, 2006.

[98] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Cross Validation Through Two-Dimensional Solution Surface for Cost-Sensitive SVM," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 6, pp. 1103-1121, 2017.

[99] S. J. Jeffrey, J. O. Carter, K. B. Moodie, and A. R. Beswick, "Using spatial interpolation to construct a comprehensive archive of Australian climate data," *Environmental Modelling & Software,* vol. 16, no. 4, pp. 309-330, 2001.

[100] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology,* vol. 49, no. 4, pp. 764-766, 2013.

[101] G. Strbac, C. K. Gan, M. Aunedi, V. Stanojevic, P. Djapic, J. Dejvises, P. Mancarella, A. Hawkes, and D. Pudjianto. "Benefits of Advanced Smart Metering for Demand Response based Control of Distribution Networks," 2010, https://www.energynetworks.org/industry-hub/resource-library/smart-meters-benefits-for-demand-response-based-control-of-distribution-networks.pdf.

[102] M. Xu, R. Li, and F. Li, "Phase Identification With Incomplete Data," *IEEE Transactions on Smart Grid,* vol. 9, no. 4, pp. 2777-2785, 2018.

[103] H. Pezeshki, and P. J. Wolfs, "Consumer phase identification in a three phase unbalanced LV distribution network," in 2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), 2012, pp. 1-7.

[104] D. H. Wolpert, "The Lack of A Priori Distinctions Between Learning Algorithms," *Neural Computation,* vol. 8, no. 7, pp. 1341-1390, 1996.

[105] D. D. Lee, and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," in Advances in Neural Information Processing Systems 13 (NIPS 2000), 2000.

[106] O. Berné, C. Joblin, Y. Deville, J. D. Smith, M. Rapacioli, J. P. Bernard, J. Thomas, W. Reach, and A. Abergel, "Analysis of the emission of very small dust particles from Spitzer spectro-imagery data using blind signal separation

methods," *Astronomy & Astrophysics,* vol. 469, pp. 575-586, 2007.

[107] L. Fang, K. Ma, and X. Zhang, "A Statistical Approach to Guide Phase Swapping for Data-Scarce Low Voltage Networks," *IEEE Transactions on Power Systems,* vol. 35, no. 1, pp. 751-761, 2020.

[108] C.-C. Kuo, and Y.-T. Chao, "Energy management based on AM/FM/GIS for phase balancing application on distribution systems," *Energy Conversion and Management,* vol. 51, no. 3, pp. 485-492, 2010/03/01/, 2010.

[109] G. Grigoraş, and M. Gavrilaş, "Phase swapping of lateral branches from low-voltage distribution networks for load balancing," in 2016 International Conference and Exposition on Electrical and Power Engineering (EPE), 2016, pp. 715-718.

[110] M.-Y. Huang, C.-S. Chen, C.-H. Lin, M.-S. Kang, H.-J. Chuang, and C.-W. Huang, "Three-phase balancing of distribution feeders using immune algorithm," *IET Generation, Transmission & Distribution,* vol. 2, no. 3, pp. 383-392, 2008.

[111] C. Lin, C. Chen, H. Chuang, M. Huang, and C. Huang, "An Expert System for Three-Phase Balancing of Distribution Feeders," *IEEE Transactions on Power Systems,* vol. 23, no. 3, pp. 1488-1496, 2008.

[112] L. Chia-Hung, C. Chao-Shun, C. Hui-Jen, and H. Cheng-Yu, "Heuristic rule-based phase balancing of distribution systems by considering customer load patterns," *IEEE Transactions on Power Systems,* vol. 20, pp. 709-716, 2005.

[113] G. Mahendran, M. Sathiskumar, S. Thiruvenkadam, and L. Lakshminarasimman, "Multi-objective Unbalanced Distribution Network Reconfiguration through Hybrid Heuristic Algorithm," *Journal of Electrical Engineering & Technology,* vol. 8, pp. 211-222, 2013.

[114] X. Geng, S. Gupta, and L. Xie, "Robust Look-ahead Three-phase Balancing of Uncertain Distribution Loads," in 52th Hawaii International Conference on System Sciences (HICSS-52), Hawaii 2018.

[115] I. Mendia, S. Gil-López, J. Del Ser, A. G. Bordagaray, J. G. Prado, and M. Vélez, "Optimal Phase Swapping in Low Voltage Distribution Networks Based on Smart Meter Data and Optimization Heuristics," in Harmony Search Algorithm, Singapore, 2017, pp. 283-293.

[116] G. ANTOINE, L. D. ALVARO, and G. ROUPIOZ, "LARGE SCALE PHASE

BALANCING OF LV NETWORKS USING THE AMM INFRASTRUCTURE," in Cired, Frankfurt, 2011.

[117] C. Ding, X. He, and H. D. Simon, "On the equivalence of non-negative matrix factorization and spectral clustering," in in SIAM International Conference on Data Mining, 2005.

[118] J.-P. Zimmermann, M. Evans, J. Griggs, N. King, L. Harding, P. Roberts, and C. Evan. "Household Electricity Survey A study of domestic electrical product usage," 2012, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/208097/10043_R66141HouseholdElectricitySurveyFinalReportissue4.pdf.

[119] R. Ash, T. Butler, R. Potter, and D. Hollingworth. "Opening up the smart grid," 2017, https://openlv.net/wp-content/uploads/2017/10/OpenLV-Measurement-Points-V1.0.pdf.

[120] Y. Zhang, F. Li, Z. Hu, and G. Shaddick, "Quantification of low voltage network reinforcement costs: A statistical approach," *IEEE Transactions on Power Systems,* vol. 28, no. 2, pp. 810-818, 2013.

[121] D. L. Davies, and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-1, no. 2, pp. 224-227, 1979.

[122] C. Gu, W. Yang, Y. Song, and F. Li, "Distribution Network Pricing for Uncertain Load Growth Using Fuzzy Set Theory," *IEEE Transactions on Smart Grid,* vol. 7, no. 4, pp. 1932-1940, 2016.

[123] W. H. Kersting, *Distribution System Modeling and Analysis, 4th Edition*, p.^pp. 39-77: CRC Press, taylor& Francis Group, 2017.

[124] M. Liozu Stephan, and A. Hinterhuber, "Pricing orientation, pricing capabilities, and firm performance," *Management Decision,* vol. 51, no. 3, pp. 594-614, 2013.

[125] K. S. Wolske, K. T. Gillingham, and P. W. Schultz, "Peer influence on household energy behaviours," *Nature Energy,* vol. 5, no. 3, pp. 202-212, 2020.

[126] H. Pettifor, C. Wilson, J. Axsen, W. Abrahamse, and J. Anable, "Social influence in the global diffusion of alternative fuel vehicles – A meta-analysis," *Journal of Transport Geography,* vol. 62, pp. 247-261, 2017/06/01/, 2017.